Essays in Behavioral Economics

Thesis advisor
Drew Fudenberg & Alvin Roth

Author
Alexander Peysakhovich

Essays in Behavioral Economics

Abstract

Essays in this dissertation cover three topics in behavioral economics: social preferences, ambiguity aversion and self-control.

The first essay, based on work with Aurelie Ouss, studies the behavior of individuals making decisions to punish norm violators. It addresses two types of questions. First, what parameters affect these punishment decisions? Second, what do outcomes look like when these decisions are aggregated? Experimental data show that individual punishment decisions appear to respond to individual cost and not necessarily social cost. Additionally, individuals appear not to take the probability that violators will be apprehended into account. Finally, punishment by others does not act as a perfect substitute for own punishment. These combined effects mean that aggregate levels of punishment rarely resemble those in line with commonly used benchmarks such as optimal deterrence.

The second essay, based on work with Uma Karmarkar, studies how information affects valuation of ambiguous financial prospects. Experimental results show that across several domains individual valuations appear to react much more strongly to favorable information than unfavorable information. Additional studies indicate that this effect is driven by two mechanisms. The first is a bias towards the integration of favorable information. The second is an effect of ambiguity aversion, individuals appear to be averse to subjective ignorance and so unfavorable information has a positive component: it removes some of this uncertainty.

The final essay looks at how dual-self (Fudenberg-Levine (2006)) decision makers can use commitment contracts to combat self-control problems and implement long-run optimal behavior. The main results show that both stick contracts, which levy a fine when an individual gives in to a temptation, and carrot contracts, which give rewards for resisting, can simulate binding commitments. However, carrots have several advantages over sticks. Sticks create a temptation to cancel the contract, carrots are less vulnerable to trembles and finally carrots allow for more flexibility.

Contents

Acknowledgments

While my name appears as the sole author of this thesis, any intellectual contributions made inside are not simply my own but rather a product of the fruitful collaboration, mentoring and help I have received over many years from a large group of dedicated individuals. I hope that I can continue to collaborate with them, learn from them and consider them my friends for many years.

First, I would like to thank the co-chairs of my committee, Drew Fudenberg and Al Roth. In my years at Harvard, they have been valuable mentors and without them I would not be where I am today. From Al I learned the importance of drawing questions from the real world and that lab experiments could form a valuable part of an empirical portfolio. From Drew, I learned the power of carefully done theory to illuminate the important aspects of a problem. I thank both of them for the time they spent, for financial support and for throwing awesome barbecues every spring.

Uma Karmarkar and David Laibson also served as important mentors for my time at Harvard. From Uma I learned the language of neuroscience and psychology and the importance of interdisciplinary communication. David was a constant source of upbeat inspiration and would always offer a new perspective on things whenever I brought to him a project that I was currently working on. From the two of them, I also learned the diplomacy skills required to navigate interdisciplinary projects (though this lesson may not have been completely absorbed).

I also thank my undergraduate advisers Adam Brandenburger and Ennio Stacchetti for turning me onto economics and research as a possible career. I have not regretted this choice.

I've learned much from my friends and classmates at Harvard and this thesis contains many ideas that have been generated either in academic or casual conversations with them. I also thank my lab mates at Harvard's Program for Evolutionary Dynamics and the future Human Cooperation Lab, the combination of which will become my new academic home. Finally, I thank the members of Paul Glimcher's lab at NYU, where I learned a great deal of important material that has since come to shape my thinking. I also thank the staff at the HBS CLER and KSG Decision Science Lab for all their assistance in setting up and running experiments.

I would like to thank some specific individuals who stand out from these groups: David

Baqaee, Natalie Bau, Yochai Benkler, Flo Blume, Dan Burghart, Peter Coles, Tom Cunningham, Michael Denny, Derek Dunfield, Peter Ganong, Paul Glimcher, Tom Gole, Duncan Gilchrist, Simon Jager, Stephanie Hurder, Scott Kominers, Danial Lashkari, Robert Molony, Martin Nowak, Marco Piovesan, Mikkel Plagborg-Moller, David Rand, Inna Sanamyan, Jesse Schreger, Amitai Shenhav, Dmitry Taubinsky, Ryan Webb, and Oren Ziv.

I thank my parents for their support and my grandfather, who spent a great deal of my high school career nagging me about the importance of taking mathematics seriously. I guess it was a good idea.

Finally, I wish to thank Aurelie Ouss. You have supported me when things were bad and celebrated with me when things were good. You've inspired me to do better when I did not think I could. I've learned many things from you and I would not be where I am without your love and support.

Chapter 1

Introduction

> "The great enemy of the truth is very often not the lie - deliberate, contrived, and dishonest - but the myth - persistent, persuasive, and unrealistic."
> John F. Kennedy

When I was an undergraduate in university, I had a good friend with whom I would cook dinners occasionally. It was college level cooking, so it generally took the form of pasta and tomato sauce. One time when we were preparing one of these dinners, my friend went to pour hot water from the tap into the pasta pot. I asked what he was doing and told him that he should use cold water instead because it would boil faster than the hot water. He didn't believe me at first but after I explained that this was because molecules in the cold water were generally closer together and thus could more readily transmit energy to each other, he agreed. He changed his action due to a seemingly plausible theory (which also happened to contradict all laws of thermodynamics). I must have thought that the incident was funny but I forgot about it afterwards.

Five years later, when I was in graduate school and he was about to finish his J.D. and start a law job in New York we met for drinks. He asked me if I remembered what I told him about cold water and pasta; apparently, he'd spent the last 5 years of pasta making following my advice and only recently found out that my theory was, in fact, complete nonsense. I add that this only happened because he tried explain the method to another friend, at the time employed as a chef.

I don't think this story is a result of any lack of intelligence on the part of my friend or

any idiosyncratic defects in his way of thinking. Rather, I think that there are a great deal of theories which guide our daily behaviors the assumptions of which we take for granted as true and the conclusions and prescriptions of which we view as a guide for our decisions. The intellectually easy shortcut of looking at our favored theories as complete pictures of the world, rather than at best hastily sketched maps, unfortunately often substitutes for a more nuanced way of thinking and often leads to colossal mistakes. It would be nice to think that scientists are safe from this bias, but that appears not to be the case. Indeed, a large source of impasse in the social sciences has historically been the proliferation of strongly held theories without correspondingly strong empirical evidence of their relevance to the world. Luckily, in economics, this is beginning to change.

One trend facilitating this change is the free exchange of ideas between academic fields, for example the blending of insights from economics, psychology, biology and neuroscience. My own work fits into this category and my general interest is in taking psychology's ability to pry into individual behavior and integrating it into economics to understand how individual behaviors turn into aggregate outcomes. There are many methodological debates about the relevance of psychological and neuroscientific variables to economists. I leave these debates, as well as those about what constitutes 'real economics,' to others and instead offer the following collection of essays as empirical evidence to the value of combining methodologies, insights and ideas from different fields. I hope the reader agrees.

The first essay, based on work with Aurelie Ouss, deals with punishment decisions made by individuals, such as those implicitly made when individuals vote for legislators who set laws, judges who enforce them or serve on juries to render verdicts and sentences. We consider the simple case of an individual who has a choice to sanction another individual who has just committed a norm violation. We ask two sets of questions. First, what parameters of the decision problem affect this individuals decision? Second, if we consider particular social benchmarks as target outcomes (eg. those given by optimal deterrence), will the aggregation of many individuals behaviors reach, undershoot or overshoot these benchmarks? The essay includes both a simple economic model of this behavior and experimental explorations of its implications. We find that punisher behavior cares more about private costs than social costs, ignores probability of capture and does not seem to be crowded out by other punishers. Thus, we argue, that hopes of reaching levels of punish-

ment consistent optimally deterrence in institutions driven by individual decision-makers are slim.

The second essay, based on work with Uma Karmarkar, discusses how individuals form evaluations of financial prospects when they only have partial information about the probabilities of different outcomes, i.e. ambiguity. We find that favorable information appears to affect individual decisions much more than unfavorable information, contrary to a standard Bayesian explanation. Additionally, we find that two psychological mechanisms appear to drive this behavioral asymmetry. First, there is a general tendency for individuals to overweight positive information in their decisions, an effect that looks quite similar to what has been called confirmation bias in the psychology literature. Second, we find that a general aversion to subjective uncertainty is at play - two factors appear to play a role in individual evaluations of ambiguous gambles: individuals care both about their reported estimate of a favorable outcome but also in how certain they feel about the estimate. In this case, favorable information has two positive components: it increases estimates of favorable outcomes and increases certainty. However, unfavorable information has one negative component (decreasing estimates) but also one positive one (increasing certainty). This asymmetry in effects then creates a behavioral asymmetry in evaluations. Since most situations include an element of ambiguity, I hope that this work becomes a part of a solid empirical bedrock upon which to build better, simpler and more useful theories of decision-making in the presence of partial ambiguity.

The topic of the final essay is commitment and self-control. We often do things that we know we shouldn't: we plan to work, but we procrastinate, we plan to start eating healthy, but only after today's desserts and we plan to start spending more responsible, but first, we buy a new useless gadget. For economists, this behavior is puzzling because it suggests a fundamental departure from a rational, time consistent agent. For the rest of us, this behavior is normal but begs an important question: what can we do about it? My final essay applies an economic model of self-control to the second question. The model consider self-control as a conflict between a short-run self who tries to maximize a discounted utility function and a long-run self who discounts at a slower rate but must use a costly resource, self-control, to impose his preferences on decisions. I ask whether particular kinds of economic contracts can help individuals facing temptation. I look at stick contracts, which

allow an individual to set a fine for themselves in the case of future bad behavior and carrot contracts which allow individuals to borrow future consumption to reward their good behavior. I show that both types of contracts can be helpful for individuals facing temptation problems but that carrot contracts have several important advantages to stick contracts. Most existing research in the psychology and behavioral economics of self-control has focused on stick contracts and binding commitments but this final essay indicates that the realm of carrots deserves a look.

Real behaviors, are necessarily much more complicated than the models and experiments these essays contain but I hope that their simplicity is a virtue. Good models and good experiments alike strip out the irrelevant complexities of a problem and focus on the important moving parts. Whether my essays are successful in doing so and in addressing important questions remains an empirical question for the reader.

Chapter 2

Cold Glow and Punishment Decisions

> "[The] Eighth Amendment is our insulation from our baser selves [when] a cry is heard that morality requires vengeance to evidence society's abhorrence of the act." - Thurgood Marshall

2.1 Brief Summary

When will the aggregation of individual punishing behaviors lead to outcomes in line with those resulting from instrumental uses of sanctions? We present a model where individuals derive private utility from punishing norm-breakers ("cold glow"), and compare their choices to those made if penalties are only viewed as a means for social cooperation. Our theory predicts that cold glow punishers take into account their private share of the cost, care about their own contribution to overall punishment, and underweight the role of probability of capture. Instrumental punishers seeking optimal deterrence care about social costs and benefits of punishments, probability of apprehension, and total levels of punishment. This means that that different environments can predictably result in either over-punishment or under-punishment relative to the benchmark of optimal deterrence. We confirm this in a series of experiments.

2.2 Introduction

Do individuals choosing punishments act in ways that are compatible with optimal levels of punishment to achieve social cooperation at minimal cost? If individuals derive private benefits from punishing norm-breakers, they will respond to different parameters than punishers only interested in maximizing material social welfare (optimal deterrence), and aggregate outcomes might differ radically. We build a theory of punishment decisions built on psychologically defensible assumptions, use it to focus on a set of parameters to consider, and test responses to variations in these parameters in a series of experiments.

There are many reasons for devoting resources towards sanctioning law breakers. For example, deterrence theory posits that higher potential punishments reduce law-breaking in a society, thus helping to maintain social cooperation. On the other hand, retributive theories see punishment as an end in itself. Motives such as deterrence could also be called 'public goods' motives of punishment, where the public good is increased social payoffs from increased cooperation; while motives such as retribution could also be called 'private goods' motives, since individuals receive personal utility from the punishment itself. Our main intuition is simple: if individuals view punishment more like a private good, then aggregations of individual decisions may not lead to outcomes in line with optimal punishments with a benchmark such as optimal deterrence in mind.[1]

We first formalize our intuition of punishment as a private good using a simple model. Our model builds strong reciprocity (see Gintis et al. (2005) for a survey) into a utility function. When an individual commits a socially praise-worthy act, that individual's payoff positively enters into the utility of others: they gain private benefits from increasing his welfare. Conversely, when an individual commits a norm violation, the utility of the norm violator negatively enters the utility of other individuals: individuals gain private benefits when a norm-breaker's material welfare is reduced. We focus on the utility from punishment aspect of our model, which we term "cold glow."[2] We compare these decisions to

[1]This benchmark, formalized by Becker (1968), is the most frequent model used in the economics of crime literature. We later discuss other possible benchmarks motivated by the idea that psychic costs and benefits can be allowed to enter the cost-benefit calculation.

[2]In reference to warm glow theories of altruism, described in Andreoni (1990) and related works.

those chosen by a Beckerian punisher interested in total social material payoffs. Compared to this benchmark, cold glow punishers respond to personal shares of cost of punishment and not to the total burden to the public, can be relatively insensitive to probability of apprehension, and punishment by others may not substitute for own punishments perfectly. These effects can lead to over or under punishment relative to the deterrence benchmark, depending on the structure of the environment.

Individual punishment decisions are important in many contexts, ranging from daily interactions to business organizations. We discuss one such domain of application: how individual decisions can shape aggregate outcomes in the criminal justice system. Our theory is most applicable to two such channels: voter behavior (the elections of judges and legislators) and juries (citizen juries set penalties in tort cases). We survey existing evidence on behavior of voters, judges and tort juries that is consistent with our theory of cold glow punishments. We then turn to another method for testing this theory: a series of laboratory experiments. Our experimental designs allow for transparent calculation of levels of punishment that would reach normative benchmarks.[3] This allows us to not only ask whether individual behavior responds to particular parameter changes but also whether aggregate behavior is, in some sense, 'optimal.'

We present three experiments, in which people can punish a norm violation: taking from a third party. We vary conditions of sanctions in order to test the role of different parameters in punishment choices, and how individual behaviors aggregate up.

Our first experiment looks at how punishment choices respond costs. The punishment available in this experiment is excluding norm breakers from the game: when this happens, they can neither make money nor take from other players. We show that environments where individuals can punish norm-breakers but do not personally bear the full cost of their decisions can lead to socially inefficient over-punishment. Our setup is such that relatively small punishments can implement social goals consistent with motives of general deterrence, specific deterrence and incapacitation; yet when costs are not fully internalized,

[3]Many papers consider the addition of punishment to public goods games (Ostrom et al. (1992)), dictator games (Fehr and Fischbacher (2004)). Others ask for individuals' impressions of 'fair punishments' in survey scenarios (Baron and Ritov (1993), Sunstein et al. (2000)). However in these games the calculations for material payoff maximizing punishments are not as transparent.

players over-punish. Results from this experiment allow us to conclusively rule out 'public goods' motivations as the sole drivers of high levels of punishment.

Our second experiment investigates the role of probability of apprehension in punishment choices. A player can take from a third party, and we experimentally vary the probability with which he is found (high or low). We compare ex-ante punishment choices and taking behavior across conditions. Consistent with our theory, choices of penalty do not react to changes in probability of apprehension, but taker behavior does. This leads to a different kind of inefficient punishment: levels too low to deter socially destructive behavior. We replicate these results in a third experiment where assigners give penalties as a reaction to decider behavior and not as an ex-ante deterrent.

Our final experiment looks at whether our 'cold glow' terminology is apt. The theory of 'warm glow' (Andreoni (1990)) posits that individuals gain private benefits from the *act* of contributing to a public good and not from the total share provided. In our final experiment, we ask whether individuals gain private benefit from overall levels of punishments imposed on norm-breakers, or whether these psychic benefits come from *their own* contributions to the punishment. In our study, two individuals make punishment decisions in sequence. We look at whether the second decision-maker's punishment decreases with the punishment of the first individual, and find that on average, no crowd-out occurs. We replicate these effects in an experiment where the first punisher's decision is made by a computer.

The rest of our paper is organized as follows: in the next section we present our model. In section 3, we review empirical evidence on field behaviors consistent with cold glow. Sections $4-6$ present our experiments, which examine respectively the impact of cost structures, the effects of probability of apprehension, and crowding out. Section 7 concludes.

2.3 A Model of Third-Party Negative Reciprocity

We first present a model punishing behaviors. We look at aggregate outcomes when punishers care about material social payoffs, and when they also derive utility from punishment itself. We focus on third-party punishment, and ask what parameters affect decisions

depending on the punisher's motivations.

2.3.1 General Setup

We begin with a simple game with three players: a taker (T), who can take or not take $\{t, nt\}$ from a victim, (V); and a punisher (P) who chooses s_P, how much to sanction the taker. If T chooses to take, V loses $s_T > 0$, and T gains αs_T. An individual who chose t is caught with probability p, in which case, they receive the sanction chosen by P, s_P. We treat V as a passive observer. The taker and the victim's utilities are given by their material payoffs:

$$\begin{aligned} U_T(s_T, s_P) &= \alpha s_T - s_P \\ U_V(s_T) &= -s_T \end{aligned}$$

We have $\alpha \in [0, 1]$, so that taking is a socially destructive action. Finally, we assume that sanction level s_P has a cost of $\beta > 0$ per unit paid for by the punisher P.

We will consider punishers with two types of social preferences: first, a punisher who cares about total material welfare; second, a punisher who gains personal utility from punishing socially destructive actions. We will apply these models to three types of sanctions: ex-post punishment, ex-ante punishment commitments, and punishment in the presence of several punishers.

2.3.2 Material Social Payoff Maximizing Punishers

We first consider the case of a punisher who cares about total material welfare: his goal is to minimize harm, subject to cost. While we allow for flexibility in assessment of harm and in the relative weight of pro-social and individual considerations, the punisher's problem is similar to that of the social planner in Becker (1968), so we will refer to him as a Beckerian punisher.

The Beckerian punisher's utility from the action pair (s_P, s_T) is given by

$$U_P(s_P, s_T) = -\beta s_P + \gamma(U_V(s_P, s_T) + \phi(s_T)U_T(s_P, s_T)).$$

The first term reflects P's material payoff, which can only be affected by his choice of sanction. The second term is P's social preference; γ measures how much weight P puts on maximizing social efficiency relative to his own payoffs. So $\gamma = 0$ represents a standard self-interested actor, and $\gamma = \infty$ represents an individual who cares only about the total material payoff of the rest of society, ignoring his own payoff.[4] We let $\phi(s_T)$ represent the weight of the taker's utility in the punisher's maximization, which can depend on T's actions. When the taker does not take, $\phi = 1$ but if they choose to take, P may put a lower value on the taker's payoff than on the victim's.

The ex-post punishment case is trivial: in a one-shot interaction, P would choose a punishment level of 0, since there are only costs and no benefits to punishment. The ex-ante case, where P commits to a publicly known level of punishment s_P before T makes their decision, is more interesting. Recall that when he chooses to take, T is found with (exogenous) probability p, in which case the sanction applies. The taker is perfectly aware of this law when making her decisions, so she chooses to take if

$$U_T(s_T, s_P) \quad = \quad \alpha s_T - p s_P \quad > \quad 0$$

From this equation it follows that there is a level $s_P^{Deter}(p)$ above which T will not take, while below which T prefers to take and that this s_P^{Deter} increases in p. For simplicity, we assume that when indifferent, T chooses not to take. To avoid off equilibrium path dynamics, we assume that T trembles to an unintended action with probability ϵ which is small.

We can also look at P's utility in different cases:

$$U(s_P) = \begin{cases} 0 & \text{if } T \text{ didn't take} \\ -s_T + \phi(s_T)(\alpha s_T) & \text{if } T \text{ took and was not found} \\ -\beta s_P - s_T + \phi(s_T)(\alpha s_T - \beta s_P) & \text{if } T \text{ took, was found and } s_P \text{ was applied} \end{cases}$$

Assuming that the taker is rational as above, the punisher can maximize this utility with backward induction. We can show that the only levels of punishment that P ever chooses

[4]The $\gamma = 1$ case, where an individual maximizes total material social payoff including his own in the welfare calculation, is most analogous in our context to the social planner Becker (1968) crime reduction function.

are 0 or $s_P^{Deter}(p)$. The punisher wishes to deter T from taking to maximize material social welfare, but if the potential costs (eg. in the case of a tremble) are too high, then this may not be worth it.

Note that this choice also depends on the level of γ, the weight that P puts on social material welfare relative to his personal costs. This gives the following important implication: if P's chosen punishment is not paid for by himself, but by a fourth player (the public) then P's maximization problem becomes exactly that of a social planner maximizing total material welfare. Thus, moving punishment costs from being private to being borne by the public can only improve total material social welfare with a Beckerian punisher. If ϵ is small, under such a publicly funded punishment scheme Beckerian punishers will always set $s_P^{Deter}(p)$, and so punishments decisions will respond strongly to variations in probability of capture.

Note also that if we think about not a single Beckerian punisher, but two identical individuals P_1 and P_2 who each set a punishments, it is easy to show that in any equilibrium of the game we will have that

$$s_{P_1} + s_{P_2} \in \{0, s_P^{Deter}(p)\}.$$

Thus, Beckerian punishers will respond to variations in total social cost, probability of capture and care about total levels of punishment. We now introduce a model of a cold glow punishment choices and study how aggregate decisions differ.

2.3.3 Choices of Punishment with Negative Reciprocity

We now assume that individuals receive private benefits from negatively affecting the payoffs of those who have done socially inappropriate actions. We call these private benefits cold glow. Though we do not present it here, our model could be expanded to allow for individuals to get a private benefit, or warm glow (Andreoni (1990)), from positively affecting the payoffs of those who have done socially appropriate actions.

We argue that our assumptions about cold glow can be justified empirically. A large literature in behavioral economics points to the fact that individuals will often sacrifice personal payoffs to reduce the payoffs of individuals who behave selfishly in games such as

the public goods game (Ostrom et al. (1992)) or the dictator game, even when the individual choosing to sanction is a third party and has no personal stake in the game itself (Fehr and Fischbacher (2004)). This occurs even when punishments can't be used to 'teach a lesson.[5][6]

Studies in social neuroscience give evidence that this behavior is driven by pleasure gained from the sanctions themselves: activity in the brain's reward areas during costless punishment can be used to predict punishment behavior in costly punishment situation (De Quervain et al. (2004)). Furthermore, individuals show reward activity (which correlates with subjective reports) when they watch another individual who cheated them in a trust game receive electric shocks, but not when the shocks are given to an individual who had been cooperative (Singer et al. (2006)).

Finally, research in moral psychology hints at the final part of our assumption, which is that this motive is very blunt: it is 'turned on' by harm itself and not by intention to harm. Cushman et al. (2009) ask individuals to play a modified dictator game, in which the dictator chooses between dice, with each different die yielding different probabilities of fair or selfish allocations. After the die is rolled, recipients are allowed to punish or reward the dictator. The authors find that outcomes predict punishment or reward behavior by the recipients, while intentions (choice of dice) have a smaller effect.

We first discuss these preferences in an ex-post decision. We then show how they affect ex-ante punishment decisions, that is, choices of punishment made when individuals can credibly commit to sanction a behavior in the future.

Ex-Post Behavior

First, we introduce a basic model of social preferences which depend on the *action* taken by another player.[7] We start with simple three player model, and then extend it to N

[5]Fudenberg and Pathak (2010) show that individuals will pay to punish others who behave anti-socially in public goods games even when the effects of the punishment are not known until the end of the session.

[6]We also note that harmful acts appear to be punished more harshly when they are caused more directly. For example, Coffman (2011) shows that third parties punish a harmful act more when an individual himself commits it than when the same individual uses an intermediary to create the same outcome. To keep our discussion simpler, we omit such motivations from our model.

[7]Theories of social preferences in economics can be divided into several categories: theories such as inequity aversion (Fehr and Schmidt (1999)) take outcomes as the objects over which utility functions are

players.

Three Players The utility functions of the Taker and the Victim are the same as in the previous sub-section. However, the Punisher's utility is now a function of both his material payoff (the first term), and his reaction to the Taker's action:

$$U_P(s_T, s_P) \quad = \quad -\beta s_P + \lambda(\Delta U_T(s_T, s_P), s_T)$$

The second term in P's utility, λ, captures the punisher's social preferences. The first argument of this function, $\Delta U_T(s_T, s_P)$ is the *total change* (relative to some baseline) to the taker's utility that occurs as a result of the punisher's action. Note that because T's utility is linear in s_P, the choice of baseline doesn't matter. Since s_T is fixed from the punisher's perspective when the punishment is carried out, we simplify the arguments to $\lambda(s_P, s_T)$.

The second argument, tells us how T's actions affect P's social preferences over T's payoff.

We make the following assumptions:

Assumption 1. *λ is smooth and concave in s_P.*

This is a standard assumption so that we can use our tools of maximization. Note that we do not very much constrain the shape of λ. In particular, for any s_T, λ can either be always increasing (bigger punishments are always better) or reach a global maximum for a certain value of s_P, which can be thought of as the 'perfectly fair' punishment, in line with just desserts theories.

Assumption 2. *We have that* $\dfrac{\partial^2 \lambda(\cdot, s_T)}{\partial s_P \partial s_T} > 0.$

Assumption 2 is the driving assumption of our model. It states that as the taker's action becomes more inappropriate, the punisher's attitude towards T's payoffs becomes increasingly negative (recall that higher s_T means a larger transfer from V when T chooses to take). Our final assumption is a normalization:

---

defined, while fairness theories (eg. Rabin (1993)) take intentions as the important objects. By contrast, we take *actions* as well as payoffs as the primary focus of our theory, in this way we are similar to social norms theories such as Axelrod (1986)).

Assumption 3. *We have that* $\frac{\partial \lambda}{\partial s_P} = 0$ *if* $s_T = 0$.

This tells us that $s_T = 0$ is a 'neutral' action which causes P to not feel either positive or negative strong reciprocity towards T. Note that in a generalized model we could relax smoothness assumptions for λ and our main results would hold. We maintain these assumptions to make exposition easier.

Note that because s_T is fixed for ex-post behavior, *levels* of λ in s_T are irrelevant for predicting P's static behavior. However, assumptions on this will make important statements about dynamic behavior or welfare. In particular, this allows for the both the situation where $\lambda(0,0) \geq \lambda(s_P, s_T \neq 0)$ implying that P would prefer to be in a situation where T takes a neutral action than where he has to exercise reciprocity or the opposite. We turn to this discussion later. However, without any assumptions on this we can still characterize behavior for a given s_T:

Proposition 1. *For any* β, s_T *there exists an optimal action for the punisher* $s^*(\beta, s_T)$. *Moreover*

1. $s_P^*(\beta, s_T)$ *decreases in* β.
2. $s_P^*(\beta, s_T)$ *increases in* s_T.

The comparative statics are easy to see: first, as the price of transfers (β) increases, P will provide less of it. Second, P's sanction of T is increasing in the inappropriateness of her behavior. There is already some evidence that punishment responds to both prices and inappropriateness in these ways: Anderson and Putterman (2006) find that punishment in public goods games responds to price effects as a normal good[8] while Peysakhovich and Rand (2012) find that reported inappropriateness ratings correlate with punishment decisions in a dictator game with third party punishment.

We note that we do not need to make these assumptions about how T's action affects V's payoff. Our model is perfectly consistent with a scenario in which T chooses an action s_T from a continuum, those s_T being linearly ordered by 'social inappropriateness', where higher s_T actions are considered more inappropriate by P.[9]

[8] These are not exactly our scenarios as public goods game punishments are not third party.

[9] Because we take appropriateness as exogenously given, an important expansion of our project would be

Four Players We now move to the case when several individuals observe the taker's behavior and can choose to affect his payoff. Suppose now there are four players: the taker T who can choose to take from the victim V, and two punishers, P and P_2, who move sequentially and can each choose how much to punish the taker, with P_2 knowing P's decision. T's utility now looks as follows:

$$U_T(s_T, s_P, s_{P_2}) \quad = \quad \alpha s_T - s_P - s_{P_2}$$

P's utility function is now as follows:

$$U_T(s_T, s_P, s_{P_2}) \quad = \quad -\beta s_P + \lambda(s_P, s_T, s_{P_2})$$

Keeping the old assumptions on the shape of the λ function, there exists a unique $s_P^*(\beta, s_T, s_{-b})$ describing the original punisher's optimal choice. What this setup gives us relative to the three-player setup is the possibility to discuss how P's punishment choices interact with that of all other punishing agents. Specifically our model allows for several types of behaviors:

Definition 1. *We say that if:*

1. $\dfrac{\partial s_P^*}{\partial s_{P_2}} < 0\ (= -1)$*, there is (perfect) crowding out*
2. $\dfrac{\partial s_P^*}{\partial s_{P_2}} > 0\ (= 1)$*, there is (perfect) crowding in*
3. $\dfrac{\partial s_P^*}{\partial s_{P_2}} = 0$*, P's punishment choice is independent of other punishers'*

Crowding out happens if a punisher considers that his and other players' punishment choices are substitutes to some degree. When crowding out is perfect (as was the case for the Beckerian punisher), a punisher only cares about is the overall level of punishment, similar to the maximum-deterrence punisher. When crowding out is imperfect, the punisher also cares about how *he* changes the taker's utility.[10]

---

to consider how appropriateness of various actions can be endogenously generated. We leave this nuance for future work.

[10]This is the flip side of 'warm glow' as discussed in Andreoni (1993) or Cornes and Sandler (1994) for public goods contributions.

Crowding in, on the contrary, implies that P's choice of punishment will be an increasing function of the other players' choices of punishment: the more other players punish, the more P punishes. Our model is mostly reduced form; one interpretation is that crowding in results from an imperfect knowledge on P's part about of how wrong T's action was. With this uncertainty in place, other players' actions serve as a signal and so can lead to crowding in of punishments.[11]

Note, again, that we do not make any assumptions on the behavior of *levels* of λ in s_{-b}, as this variable is exogenous for P. For example, we make no assumptions about whether P prefers situations in which other individuals are also allowed to punish guilty individuals. Note that assumptions on this will also inform dynamic behavior.

Which behavior holds at individual level in an empirical question. The overall aggregate levels of punishment in society will depend on the relative proportions of decision-makers who display either behavior. We study this question in experiment 3.

Ex-Ante Punishments

So far, we have only looked at P's ex-post punishment decisions, taking T's action s_T as fixed. However, most punishment decisions are set ex-ante: laws and rules are set out and potential norm-breakers are presumed to know the laws. To better understand this situation, we now turn to incorporating ex-ante motives into our theory of punishment behavior. We do so in a highly reduced form way to get our main intuitions across.

First, we modify the order of the game and simplify the strategy space: P first sets out a sanction, s_P, to which he commits. T, having seen this sanction, makes a choice from the set $\{t, nt\}$ where t (taking) is some fixed $s_T > 0$ and NT is $s_T = 0$. If T chooses t, she gets a fixed benefit of k; she is caught and has P's sanction applied to her with probability p. Note here that P only pays for the sanction if it has to be implemented.

We assume that P has a map $\psi(s_P, p)$ which represents his probabilistic assessment that T will choose t given a sanction of size s_P. Further, we assume that the function is smooth, that $\psi(s_P, p)$ decreases in s_P, that ψ is bounded away from 0 to avoid off equilibrium dynamics and that the cross partial is negative. Intuitively, these assumptions correspond to

[11] As in Bardsley and Sausgruber (2005) or Glazer and Konrad (1996).

P believing that higher sanctions decrease taking and that higher sanctions decrease taking more when probability of being caught is higher.

We leave open many possible choices of ψ,. For example, P could have rational expectations about T's behavior. One way to create a particular choice for ψ is to assume a well behaved distribution of types k for T with $k \in [0, k_{max}]$ distributed according to pdf $f(\cdot)$. Each type gets utility k from choosing t and uses a simple cost benefit tradeoff between expected sanction and expected benefit to make decisions and trembles with probability ϵ. If P does not know T's type, but knows the distribution $f(\cdot)$, we will obtain a ψ function that satisfies our criteria. We also leave open the possibility that P may be partially strategically naive: ψ can be derived from a level-k thinking (Costa-Gomes et al. (2003)) or cognitive hierarchy (Camerer et al. (2004)) model.[12]

To make ex-ante decisions, P maximizes the following expected utility:

$$\psi(s_P, p)[p(\lambda(s_P, T) - \beta s_P) + (1-p)(\lambda(0, T))] + (1 - \psi(s_P, p))\lambda(0, 0).$$

Note that now the difference in *levels* $\delta(s_P) = \lambda(0,0) - \lambda(s_P, T)$ matters. If $\delta(s_P) > 0$ for all possible values of s_P, P prefers to be in the situations where T does not take and he does not punish than in a situation where T takes and P is forced to act. This means that P's ex-ante punishments incorporate a form of a deterrence motive.[13] Having an extra motive for punishments gives us the following result:

Proposition 2. *For generic choice of ψ there exists unique s_b^* that is the optimal ex-ante punishment. Moreover this ex-ante punishment is always weakly greater than what would be imposed for $s_T = t$ in the ex-post problem above.*

This proposition means that ex-ante and ex-post punishments are different in theory, but does not explain how large this difference is. What determines this difference is the relative

[12]We point out that understanding how accurate individuals are in their beliefs about how punishment levels affect decisions of potential criminals is an important topic at the intersection of law and psychology but we do not discuss it further here.

[13]This also means that our model nests a decision-maker who cares only about the deterrence aspects of punishments by setting $\dfrac{\partial \lambda}{\partial s_P}$ to be constantly 0. In this case $\delta(s_P)$ is exactly the weight that P puts on the social loss in payoffs that happens in T chooses t.

shapes of λ and ψ. There are three interesting cases to consider. The easiest is where P is completely strategically naive and believes that T chooses T with a fixed probability no matter the sanction. This then reduces the ex-ante decision to the ex-post punishment case.

The second case is where most of the change in ψ happens at low levels of s_P. One such example sets $\psi(s_P, p) = 1$ if $s_P < \epsilon$ and $\psi(s_P, p) = q$ for $s_P \geq \epsilon$ with q and ϵ very small. Here cold glow motives push punishments above where they would be if P simply had a taste for deterrence (which would dictate that he simply set a punishment of ϵ).

However, there could be other possibilities. Consider a scenario where $\psi(s_P, p) = 1$ for $s_P < K$ where K is large and $\psi(s_P, p) = \epsilon$ for $s_P > K$. Thus, only very large punishments are deterring, but once the threshold is reached most taking behavior goes away (this could happen, for example, if T is rational, the benefits of T are modest and p is very low). Now, add to this a $\lambda(\cdot, T)$ which is single peaked in the first argument (that is, P has an optimal 'fair' punishment) and further suppose that this peak, $\overline{s}_b$, is much smaller than K. Set β very close to 0. Now an optimally deterring punishment would be one of size K, but P may choose a punishment much lower than this. Intuitively, this is because by setting a punishment K, P commits to choosing an action that is highly suboptimal, from his point of view, in the positive probability state of the world where T takes and is punished.

Thus, while cold glow gives P a deterrence motive for punishment, it also gives him other motives which he must trade off during his decision-making. We characterize the relative sizes of some of these motives in experiment 2.

2.3.4 Welfare Implications

We now compare parameters that matter for cold glow punishers relative to Beckerian punishers, and discuss other possible social benchmarks. Let's first discuss how cold glow P chooses a punishment whose cost is shared between P and P_2 (so P pays $\frac{\beta}{2}$ per unit of punishment).

We begin with the ex-post case where T has already chosen to take and has been caught. By our analysis above, P will make the choice that equates his marginal benefit from cold glow to its marginal cost (here $\frac{\beta}{N}$). This will lead to higher levels of punishment that those chosen by the Beckerian punisher, who factors in *total* costs. Furthermore, if cold glow

is not included into aggregate welfare, or if it is a private good which only benefits the punisher, but not the rest of society, then sharing costs could lead to over-punishing. We will test this in experiment 1.

One could also assume that each member of society receives cold glow utility from punishment and has preferences identical to P, and that all choices are legitimate reflections of welfare. In this case, P acts as the representative agent for society. However, even if we take cold glow to be a legitimate source of welfare, problems can arise. For example, we can consider a simple extension to our game where individuals can select into the role of punisher.[14] With sorting in place, individuals with 'the strongest' cold glow have incentives to sort into particular positions and it is unclear that individual maximization will lead to socially optimal outcomes *even if* cold glow enters into the calculation of social welfare.

We can also consider the opposite view. Behavioral economists (e.g. Kahneman et al. (1997)) often break utility down into two components: decision utility, the maximizer of which is P's choice, and experienced utility, which can be used for welfare comparisons. Taking such a point of view, cold glow reflects how individuals make decisions but doesn't tell us the whole story about how these decisions make them better or worse off. Finally, there is the important matter of how to weigh T's decrease in payoffs against the gains of other players. Moving to the ex-ante case (for example, setting laws or voting for politicians) adds even more complications to the discussion.

So far, we've given brief and by no means exhaustive list of possible ways to think about how cold glow motives should enter into aggregate welfare calculations. However, in each of these, one thing is clear: it is quite unlikely that the solution to the individual punisher's maximization problem, or to those of many such punishers, would in general aggregate up to produce socially optimal outcomes.

Our experiments test how parameters enter into individual level decisions, and they are set up in such a way that we can calculate what punishment would satisfy the Beckerian punisher's preferences. This lets us make statements both about what individuals seem to be doing and about whether their aggregate actions lead to socially optimal outcomes, and

[14] In the criminal justice system, this could happen via matching mechanisms, for example if more punitive individuals choose to become criminal prosecutors.

if not, how badly they miss the target.

2.4 Punishment Behavior in the Field: Criminal Justice

Before we turn to our experiments, we discuss how our investigation into the interaction between psychological motives and institutional structures fits into understanding important outcomes. So far, we have developed a stylized model of punishment behavior and how this behavior can lead to situations in which punishments set by individuals miss our normative benchmark. There are many important situations in which punishment decisions can affect aggregate outcomes and where we could apply such an analysis: organizations, work in groups, driving, and so on. Here, we limit our scope to a particular application: socially provisioned punishment via the criminal justice system.

We now review existing empirical work which is consistent with our model and discuss how cold glow motivations could affect aggregate outcomes through the behaviors and preferences of voters, juries and judges. We then turn to discussing how lab experiments can be integrated into an empirical strategy for understanding the aggregate effects of individual motivations.

Demand for punishment for private motives can affect aggregate outcomes through the behavior of elected officials. First, we note that if the punishment of criminals is indeed treated by voters as a private good which is provided at public cost, this would lead to demand for punishment even in the absence of clear effects on the crime reduction. There is qualitative discussion of this phenomenon: for example, legal sociologist David Garland argues that the most publicized measures (such as three strike laws, or Megan's law) have little effect on controlling crime but tend to become law due to "their immediate ability to enact public sentiment, to provide an instant response [or] to function as a retaliatory measure" (Garland (2001)).

In addition to descriptive evidence, causal links have been identified: Berdejo and Yuchtman (2009) analyze changes in sentencing behavior of judges during election cycles. They find that judge severity increases[15] when they are close to reelection and thus un-

[15]Furthermore, the authors find that this variation is due to discretionary departure above sentencing guide-

der political pressure from constituents, and sentences fall immediately afterwards. These results cannot be explained by differential work loads due to longer sentencing and variations in the month of nomination and election allow the authors to rule out seasonality or confounding political changes. This phenomenon of pre-election increase in sentences, immediately followed by a drop, is consistent with a model in which judges' preferences differ from individual voters' decisions, which are driven by the cold glow heuristic.

Cold glow could also affect outcomes in the criminal justice system through the behavior of judges themselves. We view that as a less likely place of influence, since judges are specifically trained and make their decisions in a deliberate manner, perhaps mitigating the effects of cold glow. There has been a recent resurgence of interest in studying judicial behavior (Posner (2008), Danziger et al. (2011)) which has put forth at least some evidence that judges are subject to predictable biases, so perhaps it is not impossible that cold glow is partially at play during judicial decisions.

In addition, there is some evidence in law and economics pointing to the fact that individuals may not believe that it is "fair" to factor probability of capture into punishment decisions (see Polinsky and Shavell (2000) for a discussion and Sunstein et al. (2000) for two survey-based experiments). Insensitivity to probability of capture by punishers, an important input into optimal deterrence, is a behavior that cold glow punishers can display.

There has been no research directly assessing the effect of cost structures on demand for punishment, even though the question of costs of punishment has received attention from policy makers due to the budget crises in many states.[16] The only paper to investigates the effect of a change of costs on punishment decisions is Ater et al. (2012). They exploit a quasi-experimental change in costs of arrests in Israel: the responsibility of housing arrestees awaiting trial was transferred from local police to the prison authority. The authors find a sharp increase in arrests as a result of this policy, which is consistent with an imperfect factoring in of total costs of crime reduction when making arrest decisions.[17]

---

lines, and not greater compliance to these guidelines.

[16] In particular, in California, one response has been to transfer housing of inmates from state prisons to county jails, with the argument that this would lower overall costs of criminal justice.

[17] We note there are many other possible explanations for these results: police officers' effort provision might respond to costs, police evaluations could depend on number of arrests, etc.

Additionally, whether individual punishment decisions are crowded out by already performed punishments could play a role in labor markets. There has been discussion on the role that having a criminal record plays employability of an individual (Bushway et al. (2007), Pager (2007)). One way this can occur is through a signaling channel (Rasmusen (1996)) where conviction is a signal of poor worker. However, if cold glow motives are not crowded out by already performed punishments, there may be a second channel for this effect: a lack of hiring can act as a sanction towards an individual who has committed an inappropriate act. The relative sizes of each of these effects matter quite a bit for choices of particular policies (for example, shrouding criminal records).

All in all, a lot of empirical facts can be explained by cold glow motivations playing a role in decisions which affect important aggregate outcomes. However, these decisions are a product of many factors: elections involve many non-judicial dimensions, juries are prompted to depart from emotions,[18] and exact magnitudes of costs or probabilities of apprehension are generally not known precisely by voters, juries or judges. In order to conclusively isolate the role and magnitude of cold glow in aggregate outcomes, we would ideally need data on voter, jury and judicial behaviors responding to (quasi) experimental variations in costs of judgments and probability of apprehension. Beyond the practical difficulties of implementing such a protocol, it would be difficult even in this scenario to isolate the exact mechanisms at play. To build our understanding of how cold glow interacts with institutions, we examine the behavior of individuals in a stylized setting using a series of laboratory experiments. These experimental methods allow us to study, in a controlled environment, punishment choices which are normally hard to observe in the field. For this reason, they are an important piece of a larger portfolio of methods that can help us to analyze and evaluate how cold glow motivations affect aggregate outcomes.

[18]For example, French jurors verbally pledge that they will "not listen to hatred or malice or fear or affection; [and decide] according to [their] conscience and [their] inner conviction, with the impartiality and rigor appropriate to an honest and free man."

2.5 Experiment 1: Responses to Costs

In this first experiment we test an individual level hypothesis: when costs of punishment accrue to the group rather than to the individual, will individuals increase their punishment decisions? At the social level, the game is set up so that very low levels of punishment are sufficient to deter potential norm breakers. Simultaneously, our transfer of costs to society also increases the overall cost of the punishment beyond what would be consistent with using motives such as incapacitation as the social benchmark. We then ask: will individual punishment decisions meet or exceed our social benchmarks?

2.5.1 Experimental Design

We run a series of experiments in which we vary the availability and cost structure of sanctions. In our game, participants gain Monetary Units (MU) throughout the experiment, which are converted into dollars at a rate of 50 MU per dollar. Players are randomly matched in groups of $n = 8$ to 12 players. Each group is given a public pot of $70 * n$ MU, which is equally split amongst all members of the group at the end of the game. Each player is also individually given 30 MU at the beginning of the game.

Participants play 20 rounds (one iteration) of the following game. They are asked to solve a simple math problem, for which they receive 4 MU upon completion. They are then given the possibility to "take." If a player chooses to take, she receives 2 MU, and another randomly selected player loses 3 MU. Taking, in this case, is a socially destructive behavior; yet, in the absence of sanctions, it is a dominant strategy. When a player chooses to take, she is found out in 50% of cases. Our conditions and treatments consist of varying what happens when a player is found out.

In the "No Punishment" condition, when a player is found out, she gets a message informing her that she has been found out, but nothing more happens. In both "Punishment" conditions, when a player is found out, another random player is chosen to be her "assigner." The assigner is able to punish found out players by excluding them from the game for up to 10 rounds. We elicit punishment using the strategy method: individuals choose a punishment after making their "take" decisions and seeing whether they were taken from,

but before they are informed of whether they were found out, or if they were someone's assigner. They are asked at this point to enter an amount of penalty rounds that they would assign if they are chosen as an assigner for this round. Individuals can never be chosen as their own assigner, nor do they know which player they assign penalty rounds to. In particular, if they were taken from, there is no additional chance that they will assign a punishment to the player who took from them. In all conditions, only the assigner and the individual to whom penalty rounds are allocated learn about the punishment level chosen.

Each round of exclusion is costly, and we vary the cost structure. In the "Private Punishment" (hereafter Private) condition, if a player's punishment is chosen, they will pay 2MU from their private money for each round of punishment they have imposed. In the "Public Punishment" (hereafter Public) condition, if a player's punishment is chosen, each round costs 5 MU from the public pot. This means that in the Public condition, the private share of the cost to a particular punisher is less than 2 MU per round. This experimental setup will allow us to investigate cost effects in demand for punishment, thus determining if demand for punishment looks like demand for a private good.

As a robustness check, we include one more condition. In the "One Round Take" condition, subjects play 1 round in which they can take and punish (with the public costs structure), followed by 10 rounds in which the take option is not available. In this case, since subjects cannot take for the following rounds of the interaction, future oriented motives (incapacitation or deterrence) cannot explain any choice of punishment. This is similar to the design employed by Fudenberg and Pathak (2010) who have individuals play multiple rounds of public goods games which include sanctions

In each experimental session, individuals are first put into a group to play one iteration of the No Punishment condition. After a random rematching into new groups, they play either one iteration of Public, one iteration of Private, or 3 iterations of One Round Take.[19] We implement this design for several reasons: it allows individuals to gain experience with the experiment in the first stage, and it allows us to look for correlations between individual behavior in No Punishment and their later behavior when punishment is available.

[19]Participants are not informed about the full structure of the experiment, they are only given instructions for their current condition. However, participants are informed when the One Round Take condition is the final game in the experiment.

Our experimental design is different from other experimental designs assessing the role of non-altruistic motives for punishment. We vary the cost structure of punishment, which allows us both to discuss the institutional setup of financing sanctions, and to investigate the private benefits from punishment, using a basic economics framework. Second, the punishment in this game is not fines, as in prior experiments, but exclusion for a certain number of rounds. This is allows us to include an analysis of incapacitation, and therefore contribute to the discussion of different motives of incarceration motives in the economics of crime literature.

The experiment was conducted at the Harvard Decision Science Laboratory using the z-Tree software (Fischbacher (2007)), in June and July 2012.[20] The participants, recruited using the Decision Science Laboratory pool, were university students (mean age: 21.5 years old, 58% female) in the Boston area. We have a total of 91 participants: 39 in Public, 28 in Private and 24 in One Round Take.

Participants were given a 10 dollar show-up fee, and their experimental earnings were converted at a rate of 50 MU per dollar. The experiment took between 40 and 50 minutes to complete. Participants earned between 17 and 23 dollars. They were informed of experimental earnings for each condition independently, and their final earnings were privately announced to them at the end of the experiment.

Our main outcome variable in this series of experiments is the choice of number of rounds of punishment for potential found takers. This is our measure of how much sanction players are willing to support when facing different cost structure.

2.5.2 Theories of Punishment

There are three major normative theories of punishment in the law and economics literature: incapacitation, general deterrence and specific deterrence. Our experimental setup allow us to discuss what kind of social benchmarks each of these motives sets. We briefly present these motives and how they form benchmarks in our experimental setup. Table 2.11 in the appendix summarizes predictions for these different motives.

[20] Appendix 1 presents the experimental instructions

Incapacitation

Incapacitation is the prevention of offending by removal of offenders. Shavell (1987) determines the optimal level of punishment to achieve cost-efficient incapacitation. He finds that incapacitation to be cost-efficient, the cost of incarceration (or, in our setup, of removing a player for N rounds) has to be lower than the expected harm that individual could do while incapacitated.

In our setup, even if we assume that an individual does not respond to deterrence incentives and always chooses to take, the maximal harm that individuals can do is to take 3 MU from one (random) player in each round. In the Public condition, the cost of removing this individual is 5 MU. Thus, from a perspective of maximizing social payoffs (even if we assume that 'bad' (taking) individuals' payoffs do not enter into this calculation), the cost of incapacitation outweighs its benefits.

In the Private condition, since the cost is 2 but the social benefit is 3 there may be pro-social incapacitation motives. However, from an individual payer's perspective, the expected harm per round of a rogue individual is $\frac{3}{n}$; whereas the cost of removing the player is of $\frac{5}{n}$ per round. Thus there is no private incarceration motive either.[21]

Finally, in the One Round Take treatment, exclusion cannot be chosen for incapacitation motives, since the punishment applies to rounds in which it is impossible for the punished players to take.

Deterrence

General deterrence is the impact of the threat of future punishment on behaviors. In out setup, players cannot increase general deterrence by setting higher punishments. Only players who are found out learn about other players' punishment choices, and even then, only their assigner's choice of penalty rounds. General threats therefore cannot be emitted.

Specific deterrence, however, could be a consideration. In order to investigate this possibility, we consider several possible assumptions on takers' behaviors.

[21] One may argue that risk averse players would prefer to pay a cost of $\frac{5}{n}$ for sure rather than lose $\frac{3}{n}$ with some probability, and thus that incapacitation can be seen as a form of private insurance against rogue group members. We note that this critique does not apply in the One Round Take condition.

Assumption 1*: Takers are rational criminals.* In this case, average punishment should be ≤ 1 round. By being excluded for 1 round in 50% of cases, potential thieves lose in expectation 2 units,[22] which is exactly what they gain from taking. As long as participants taking are slightly risk averse, 1 round of punishment will be enough to deter them from taking. Excluding player for more than 1 round cannot be for only specific deterrence motives.[23]

Assumption 2 : *Takers can be "taught a lesson" if punishment is higher than a certain threshold.* We present a simple mathematical model of specific deterrence with reform in the appendix. The main results of our model is that though we can rationalize many different average levels of punishment, depending on individual beliefs, for fixed beliefs, the amount of punishment should decrease as the game gets closer to the end. This is because the value of reforming individuals decreases, since there are less rounds over which benefits from reform can be reaped, but the cost of punishment stays the same.

Assumption 3 : *Takers always take, and cannot be reformed.* This case reduces to the incapacitation case.

Finally, and regardless of our assumptions about takers' behaviors, in the One Round Take treatment, no positive exclusion can be rationalized by a specific deterrence motives, since taking is only possible in the first round of this treatment condition.

Cold Glow

Contrarily to pro-social motives, cold glow predicts that punishment in Public would be higher than in the Private. Private benefits from cold glow motives will be over consumed when costs are not fully internalized. Additionally, cold glow is the only motivation consistent with any non-zero punishment in the One Round Take condition.

---

[22]The math exercise – adding up 2 numbers – is easy: players get it right in 98,7% of cases. Furthermore, no participants systematically make mistakes: only 1 participant makes more than 2 mistakes, over the 40 additions participants are asked to do. We therefore assume that loss from exclusion for 1 round is equal to 4.

[23]One reason why individuals might choose punishments greater than 1 for specific deterrence motives is if they think that other players would punish less, because those players do not care about deterrence as a public good. There is however no reason for the average punishment in the public condition to be higher than average punishment in the private condition for this reason, unless individuals believe that others punish less in the latter compared to the former.

2.5.3 Experiment 1 Results

This first section compares Public to the Private condition. We present graphs along with body text and regression analysis in the Appendix. We then present additional evidence from One Round Take as a robustness check.

Punishment Decisions

We first look at punisher's decisions. Figure 2.1 presents the number of rounds of punishment chosen in Public and Private conditions.[24]

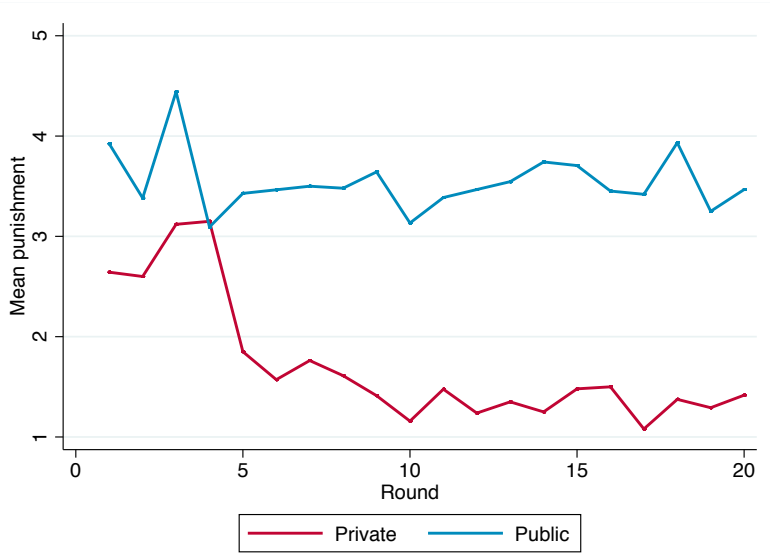

Figure 2.1: Mean punishment level chosen by round

The average begins at roughly the same level (approximately 3.5 rounds of exclusion). However, punishment decreases sharply in Private but not the Public conditions after the first 5 rounds. After this short learning period, average punishment settles to 1.3 rounds in Private and stays at 3.5 in Public.

The fact that punishment levels stay the same over rounds in the Public condition is a

[24] As a reminder: all players who are not currently excluded from the game can choose a punishment.

first indication that specific deterrence cannot be the only motivation at play: as participants get closer to the end of the game, the size of imposed punishment does not down. Furthermore, the average levels of punishment chosen in the Public condition far exceed than the levels in line with optimal deterrence or incapacitation.

Robustness check To conclusively rule out deterrence or incapacitation as the only motives for punishment, we also consider the One Round Take condition. Figure 2.2 shows the average punishment decisions made in rounds $6+$ of the Private and Public conditions, and in all iterations of the One Round Take condition. Participants in One Round Take choose an average of 2.5 rounds of exclusion compared to 3.5 rounds in Public and 1.7 in Private. The fact that One Round Take punishments are positive, and higher than in the private condition shows that cold glow, as a private benefit to punishment, is a major motivating force of punishment decisions.

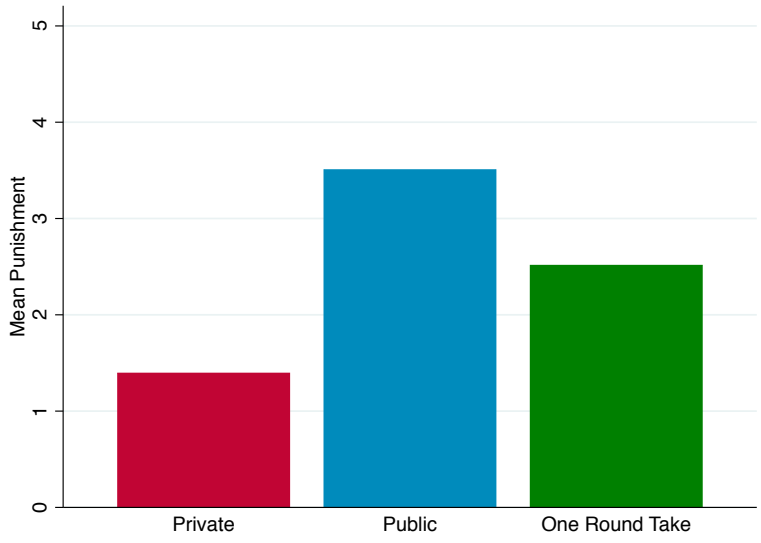

Figure 2.2: Mean punishment level chosen, round ≥ 5

Table 2.2 presents regression results that confirm the intuitions presented in the graphs. We regress amount of punishment chosen on a dummy taking values 0 for Private and 1 for Public. Standard errors are clustered by participant.

Column 1 presents results for the full sample; column 3 presents decisions made from rounds 6 to 20. Participants in the private treatment choose smaller levels of punishment than in the public treatment. This holds when we control for round effects (column 2).

Column 1 (2) of table 2.3 shows the difference in number of rounds of exclusion chosen in One Round Take and Private (Public). In this specification, the One Round Take condition is significantly higher than Private and lower than Public. In column 3, we pool the data to tease apart the relative importance of public motives (deterrence and incapacitation) and cost structures in choices of punishment. We regress punishment choices on a dummy for costs being public (Public and One Round Take conditions) vs. Private; and a dummy for public good (deterrence or incapacitation) motives (Public and Private conditions) vs. One Round Take condition. The coefficients on these dummies represent the effects of cold glow vs. public goods motives in punishment decisions. The first dummy is significantly positive: people choose more rounds of exclusion when the costs are public. The second dummy is negative, smaller in magnitude but not significant implying that non-cold glow motives play a weak role in punishment behavior in our experiment.[25]

Taken together, our regression analyses confirm that cold glow is a major motivation in punishment decisions. Other motives also exist, but cannot explain most of the variation in punishment. We now turn to see the effects of conditions on taking decisions.

Taking Decisions

Figure 2.3 shows taking decisions by availability of punishment, and table 2.1 presents our regression results. Taking behavior is significantly higher in Punishment and No Punishment conditions (column 1), which shows that general deterrence does matter: only 10% to 20% of participants who are able to take[26] choose to do so, even from round 1. However, there is no difference between the Public and Private conditions (column 3), and we find no session effects (column 2).

We find a slight learning effect in the No Punish condition. Approximately 70% of

[25] Another possible explanation for the difference in behavior between One Round Take and Public is that perhaps it is easier to ex-post rationalize punishment decisions in the former than in the latter.

[26] i.e. players who are not currently excluded from the game

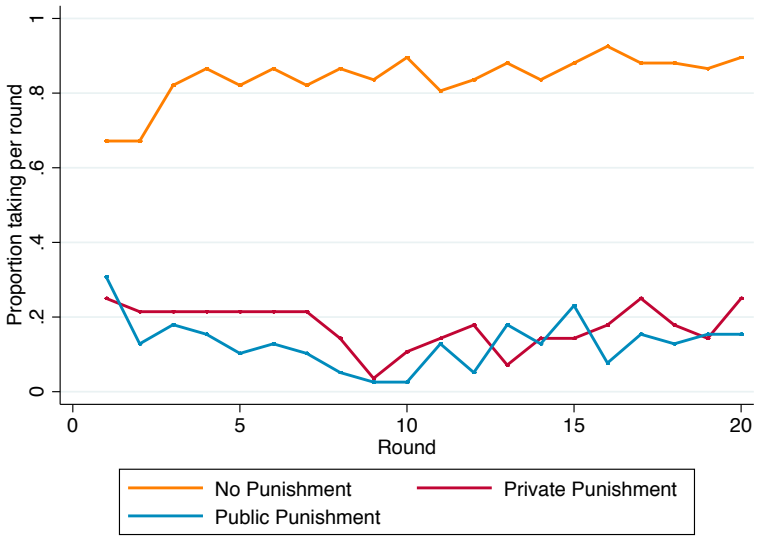

Figure 2.3: Percent choosing take

individuals take in the first round and by the 5^{th} round, 85% of participants choose to take. There is no significant difference between experimental sessions.

Individual Differences

So far, we have compared results across treatments. We now turn to individual variations within treatments. First, we ask what causes the learning effect that we find in the Private condition. We find that punishment levels decrease after a player's choice is implemented[27] in Private condition, but not in the Public condition. Table 2.4 shows the regression of punishment decisions on a dummy which takes value 1 in each round after an individual's punishment choice is implemented. On average(column 1), it appears that having paid for punishment does not influence choice of sentences (column 1). However, the effects are heterogeneous across treatment conditions (columns 2-4): in Private, subjects punish significantly less once their punishment has been chosen. We interpret this as a form of 'sticker shock.'

[27] In other words, when she is randomly chosen to be a found individual's punisher.

We also attempt to see whether behavior in No Punishment conditions predicts punishment behavior in later stages. We find an effect for individuals who take less than 15 times in the original No Punishment rounds, whom we refer to as "low takers". They also give much smaller punishments on average (Column 5 of table 2.2).[28] This result would be interesting to investigate in future experiments, as it suggests negative correlation in warm and cold glow.

2.6 Experiment 2: Responses to Probability of Apprehension

Our second experiment asks whether punishers' decisions react when probability of apprehension (and thus optimally deterring punishments) change. We see how potential norm-breakers in turn react to punishers' behaviors. If punishers don't react to these changes, this can lead to an outcome where socially wasteful low levels of punishment can occur. In addition, we compare ex-ante and ex-post punishment decisions.

2.6.1 Experimental Setup

We use a game to test both how sentences are chosen, and how potential norm-breakers respond to expected punishments.[29] The basic setup is as follows: players are matched into groups of three to play a one shot game. They begin with a balance of 80 points.

Players are randomly assigned one of three roles: assigner, taker, or target. All rules of the game are known to all players before they begin the experiment. The game proceeds as follows: the assigner commits to a publicly known level of penalty units (between 0 and 10), each of these units corresponds to a 10 point sanction. *Knowing this level of sanction*, the taker decides to take or not from the target. If the taker choose to take, they gain 20 points, and the target loses 30 points. The taker is found out with probability p. If the taker

[28] Our results are robust to changes in the definition of "low takers". We chose this specification, as it took about 5 rounds for taking behavior to plateau at 90% in the No Punishment condition.

[29] Experimental instructions are presented in the appendix

is found out, they are imposed the sanction chosen by the assigner. The assigner is charged 1 point per 5 points of sanction they assign.

Our treatments vary in the probability that the taker will be found if he takes: in the "high probability" treatment, the taker is found with a probability $9/10$; in the "low probability" treatment, with a probability $1/3$.[30] All players are informed of all rules at the beginning of the game. Final payoffs depend on choices made by all of the players. Finally, the targets make no choice in our game, but we ask them to enter what they think would be a "fair" punishment for a taker who chooses to take.

We used the online labor market Amazon's Mechanical Turk (AMT) to recruit individuals to play the game for a show-up fee of $.3$ USD and an additional payment depending on points earned, using a conversion rate of 2 points per $.01$ USD at the end of the experiment.[31]

We recruited a total of 340 individuals (mean age: 28.8, 63% male) to play this game. Each individual played exactly one role in the interaction. To make sure that all participants understood the experiment they were first given a set of instructions followed by a three question comprehension quiz (see Appendix). If they failed to answer any of the quiz questions correctly, they were not allowed to play the game. Thus all of our results are from participants who answered all comprehension questions correctly. Dropping non-comprehenders, we are left with 243 individuals (a 71 $\%$ pass rate).

---

[30] Some studies in psychology have investigated the effects of probability of apprehension on punishment decisions. These studies directly ask participants to compare hypothetical punishments in different scenarios when probabilities of apprehension change (Baron and Ritov (2009)), or asked participants to assess the relative importance of deterrence or moral motives on punishment decisions (Carlsmith et al. (2002)). In these hypothetical contexts, players state that do not want to change behaviors based on probabilities of apprehension. Our experiment adds to this literature as a very strong test of whether punishers respond to probability and deterrence motives. In our games rules are perfectly transparent and deterring punishments are very easy to calculate.

[31] Several recent studies have been undertaken to examine the validity of experimental data collected using AMT at stakes of $\sim$ 1 USD. They find that behavior on AMT matches well with standard laboratory results on economics games (Amir et al. (2012)) (Rand et al. (in Press)), and are based on samples that are more representative of the general population (Horton et al. (2011), Paolacci et al. (2010)).

2.6.2 Experiment 2: Results

Punisher Behavior

We now consider the behavior of punishers across conditions. Part a of figure 2.4 presents assigners' average punishment levels for each of the probability conditions. Mean punishment levels are exactly the same in both treatments: probability of apprehension is not a parameter individuals respond to in punishment choices. The mean punishment level is 4.0 in the high probability condition and 4.1 in the low probability, and the difference non-significant (see table 2.5).

Figure 2.4: Experiment 2: punishers' and takers' behaviors

Decisions to Take

We find that takers' behaviors, however, *do* respond to probability of apprehension on the intensive margin. We use the strategy method to elicit choices of taking: takers are asked to enter their *maximum acceptable possible penalty* (MAPP). This is a number of penalty units such that if the assigner chooses a penalty below or equal to this level, the

taker prefers to take. If the assigner chooses a larger penalty, the taker would prefer not to take. We perform analyses on choices of MAPP to understand takers' behaviors.

We first find that a relatively large amount of participants (approximately 30 %) who choose a MAPP of 0, indicating that they do not wish to take under any circumstances, in both conditions. Table 2.6 shows our regression results confirming there is no significant extensive margin response. However, focusing on the 70% of individuals who entered a MAPP > 0, we find that there is an effect on the intensive margin: as shown in part b of figure 2.4, individuals who choose to take at all choose different levels of MAPP between probability conditions (mean MAPP in low = 5.1, and mean MAPP in high = 3.8). Table 2.6 shows our regression results, confirming there is a significant intensive margin response.[32] Unlike punishers, takers respond to the probability of being caught,[33] and so the punishment levels chosen are too low to deter a lot of taker in the low probability condition.

2.6.3 Control Study: Ex-Post Punishments

A key part of our theory is that we allow for both an ex-ante (simulating a strategic motive such as deterrence) and an ex-post (or 'just desserts') component. To assess the size of these components, we ran a control experiment on AMT (n=194, age=28.9, 63 % male). The setup of the game in our control study is identical, except that the order of moves is switched: takers first choose to take or not, and then assigners choose ex-post penalties to assign to takers who are caught. We use the same probability conditions in this study. This has the added benefit of acting as a robustness check on taker behavior from our original study where one possible confound is that takers could have found the strategy method confusing.

Figure 2.5 shows the results. We find that punishers again do not respond to probability of apprehension when choosing levels of ex-post punishment (mean punishment in low = 3.4, mean punishment in high = 3.2). Takers, however, do take probability into account:

[32]We also find a gender effect. Women are less likely to take, and if they are willing to take, they enter lower maximum acceptable punishment levels. We note that this can be explained by higher risk aversion (Eckel and Grossman (2008)).

[33]This also allows us to control away a lack of attention or understanding by participants as the result of the null effect on punishment decisions as individuals are randomly assigned into roles.

25 % of individuals take in high probability condition and 43 % take in the low probability condition[34].

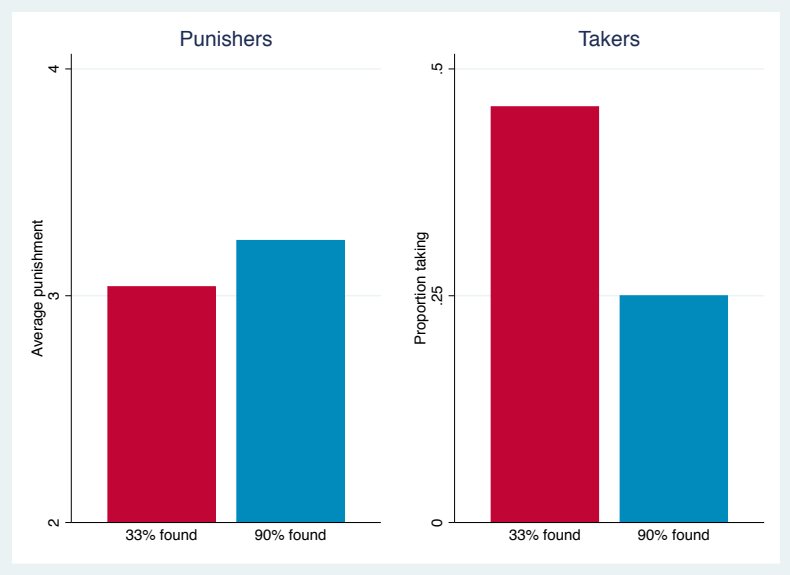

Figure 2.5: Study 2 Control Decisions

Comparisons

We now pool our data and compare the ex-ante and ex-post punishment conditions using regressions. Table 2.7 presents full sample results: as in the main sample, we find that neither choice to punish nor punishment level respond to probability of apprehension.

Furthermore, in the control condition, assigners still choose a positive level of punishment, even though this is a one-time interaction and punishments are privately costly. However, we confirm that levels of punishment are smaller when no deterrence motive is possible than when the assigner plays first: this indicates that some difference (approximately 20 percent) between ex-ante and ex-post punishments does seem to exist, however these differences are not significant. These results are consistent with the differences found

[34]This difference is significant, though only at the 10% level, due to sample size. The magnitude stays the same – 20 percentage points difference – and becomes significant at the 5% level when we control for gender

in our first experiment between the One Round Take condition and the Public condition. We conclude that some form of deterrence motives *do* exist in the punishment choices, but ex-post 'just desserts' thinking seems to be the dominant motivator of punishment behavior in our samples.

2.6.4 Fairness Judgments

Finally, we look at judgments of 'fair punishments' for caught takers from the point of view of the target. Their answers do not appear to differ across conditions (mean fair punishment in low, ex-ante = 4.3, high, ex-ante = 5, low, ex-post = 5.3, high, ex-post = 5.5).

Table 2.8 presents our regression analysis. Unsurprisingly, targets want higher punishments than assigners: this could be driven either by differences between second-party and third-party punishment (Fehr and Fischbacher (2004)), or because targets do not have to pay for chosen punishments. Interestingly, neither order of punishment assignment nor probability of being caught changes targets' beliefs about fairness: no extra retribution is demanded when probability of apprehension is lower: differences are not significant, and if anything the point estimates go in the wrong direction. All data taken together, neither punishers nor victims respond to probability of apprehension when choosing punishment levels, although this parameter seems to matter a lot in the decisions of potential norm-breakers.

2.7 Experiment 3: Crowding Out

Our final experiment asks an individual level question motivated by our theory: to what extent is punishment by one individual crowded out by known punishment choices of another individual? Our social level question asks whether a lack of crowding out can push aggregate punishment levels above particular benchmarks.

2.7.1 Main Experiment

In order to answer this question, we ran an experiment on AMT using a sample 476 individuals (mean age = 29.7, 56% male). Participants received a show-up fee of .5 USD

and an additional payment depending on their earnings during the game, using a conversion rate of 1 points per $.01$ USD.[35]

We use a game similar to experiment 2 to explore crowding out behavior. Players are randomly assigned to groups of four and start the game with 100 points. Each individual is assigned one role: assigner 1, taker, target, or assigner 2.[36] All rules of the game are known to all players before they begin the experiment. Players act sequentially as follows: assigner 1 commits to a publicly known level of penalty units $(0-6)$, each penalty unit corresponds to a 10 point sanction. *Knowing this level of penalty*, the taker decides to take or not from the target. If the taker choose to take, they gain 30 points, and the target loses 40 points. The taker is found out in $3/4$ cases. If the taker is found out, assigner 2 sees the punishment that assigner 1 chose, and is given a choice to assign an additional number of penalty units (up to 6). A found out taker is imposed the sum of the penalty units chosen by the assigner 1 and assigner 2 and both assigners are charged 1 point per 10 points of sanction they assign.

Again, although the target makes no choice in our game, we ask them to enter what they think would be a "fair" punishment for a taker who chooses to take. As in experiment 2, individuals see the instructions for the experiment and then take a quiz about the rules. Individuals who do not answer quiz questions correctly are not allowed to participate in the experiment. Overall, approximately 70% of participants answered the quiz questions correctly leaving us with 73 groups of four players.

Our main variable of interest is assigner 2's choice in level of punishment. As in the previous experiment, we use the strategy method to elicit this preference. Figure 2.6 presents the average punishment choice of assigner 2, for each possible assigner 1 choices. On average, there is no difference across assigner 1's choices, and thus no evidence of crowd-out behavior on aggregate.

We do find considerable heterogeneity in individual behavior. Because we use the strategy method we can look for different behavioral types in our population. Overall, we find

[35]Given the average completion time of our experiment and average bonuses, total payoffs amounted to an hourly wage of approximately $\$8-10$ per hour.

[36]In experimental instructions taker and target are referred to as player 1 and player 2 respectively.

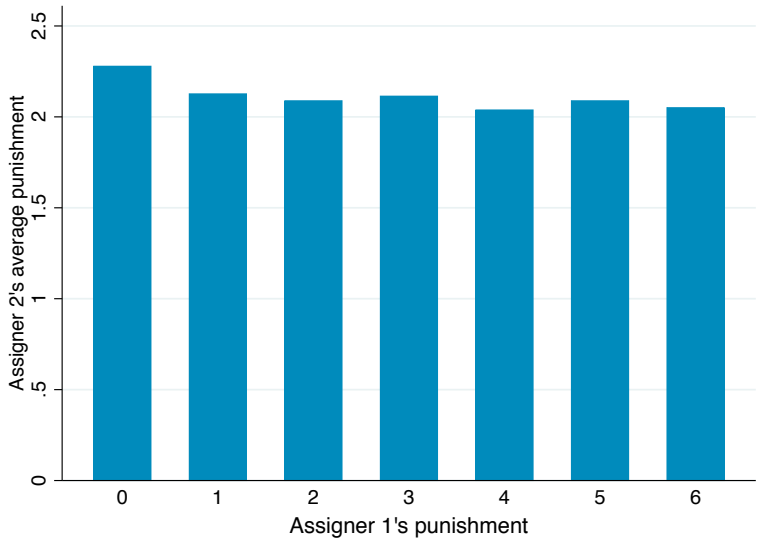

Figure 2.6: Experiment 3: Assigner 2's behavior

that approximately 80% of assigner 2's can be classified into one of three types: individuals whose sanction choices decrease in assigner 1's choice (partial crowd-out types 35%), individuals whose sanction choices increases in assigner 1's choice (crowd-in types[37] 25%) and individuals whose sanctions do not change as a function of assigner 1's choice (constant types, 20%). Individual heterogeneity is not the main focus of this discussion, so we leave as an avenue for future work. However, we can use this analysis as a robustness check. If we restrict our analysis to the crowd-out types, we still see an imperfect crowding out of own punishment by the punishment of another and we can statistically reject the hypothesis of perfect crowding out even in this restricted subsample (table 2.9).

We can also look at the average behavior of the first assigner in this experiment and what the target deems to be a fair punishment. We find that the mean punishment assigned by the first assigner is 3.02 units (30 points). Combining this with the conditional punishments of assigner 2, we find that the average total punishment on a taking player is approximately 5 units of punishment, or 50 points. We note that this is 25% higher than the mean 'fair

[37]These individuals may be using assigner 1's decision as a signal of the inappropriateness of taking.

punishment' as viewed by the targets (mean fair punishment = 42 points).

2.7.2 Control Experiment

Experiment 3 uses a strategy method and a within subject design to look for the extent of crowd-out in punishment. We ran a second study as a robustness check using a between-subject design without the strategy method. We used AMT to recruit subjects, again dropping those who failed a comprehension quiz. We were left with 243 participants (mean age = 29, 57 % male) between two conditions.

In our control experiment, players are put into groups of three and assigned a role: taker, target or assigner. All rules of the game are known to all players before they begin the experiment. The game proceeds as follows: the taker decides to take or not from the target. If the taker chose to take, they gain 30 points, and the target loses 40 points. The taker is found out in $3/4$ cases. If the taker is found out, they automatically lose c points, where c is varied to be 0 or 40 by condition. If the taker is found out, the assigner can assign up to 6 penalty units, each of which amounts to a 10 point sanction. The assigner is charged 2 point for every 1 penalty unity.

This control lets us look at crowd-out effects when punishment is assigned by an outside figure instead of another player in the game. Figure 2.7 shows the average chosen levels of punishments in the two conditions. Assigner punishment levels chosen are slightly lower when $c = 4$ than when $c = 0$, but this difference is not statistically significant, and it is in any case much smaller than a one-for-one crowding out: punishments are of on average 2 units in the $c = 0$ condition, and 1.7 in the $c = 40$ condition. Thus realized sanction are approximately 20 points in the $c = 0$ condition and 57 points in the $c = 4$ condition.

In the interest of space, we skip discussion of taker behavior and fairness evaluations by the target, as they only replicate the qualitative results of experiments 1 and 2.

This last set of experiments therefore indicates that punishment is not crowded out one for one by pre-set levels of sanctions. On average, there is no effect of pre-set sanctions on average punishment. We note that there is considerable heterogeneity in this behavior, but never observe perfect crowding out.

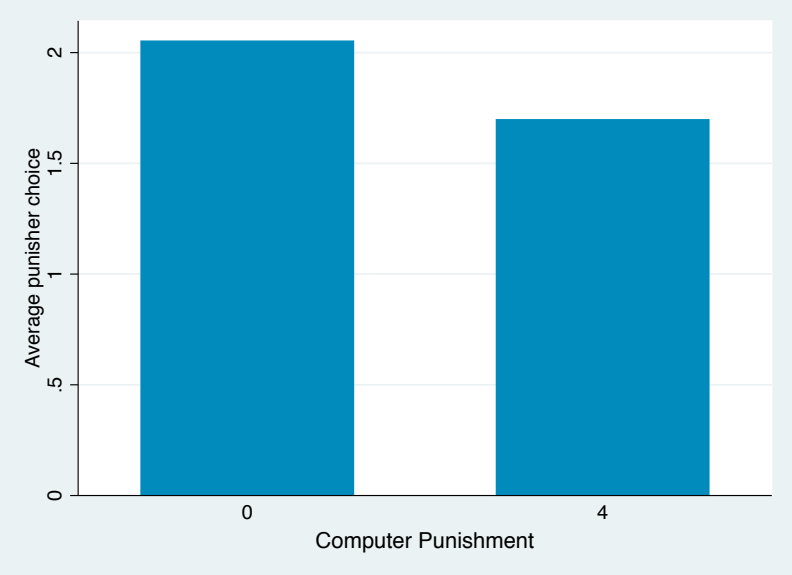

Figure 2.7: Study 2 Control Decisions

2.8 Conclusion

Though many legal scholars and philosophers think of moral reasoning as driven by rational, calculating processes, the nascent field of moral psychology suggests that moral behaviors, including the punishment of those who break social norms, are mostly driven by emotional reactions which are then rationalized by conscious processing (Greene and Haidt (2002), Haidt (2001)). Using such a blunt psychological mechanism motivated by affective factors and not rational reasoning to make punishment decisions may sometimes collaterally result in social harmony, but in other domains can result in either highly inefficient over punishing or inefficient under punishing. We have presented a simple theory based on this observation which predicts that punishment decisions will be driven by personal cost, and not public cost, will not respond to probability of apprehension, as optimal deterrence might and may not necessarily crowd-out one-for-one as punishment might in a theory of 'just desserts.' We confirm these predictions in our experiments and find little evidence that standard rational motives (deterrence, incapacitation) are major drivers of individual punishment decisions.

We argue that understanding the role more emotional or automatic mechanisms at play in choosing levels of punishments could be important in our understanding of many types of social behaviors including aggregate outcomes in the criminal justice system. We have presented several possible channels through which we believe our theory of behavior can affect these aggregate outcomes. More empirical research is needed in understanding to what extent cold glow motives drive the behaviors of voters, judges and juries, as well as everyday punishment behaviors in social groups.

Simultaneously with field data, further lab experiments could be used to investigate the mechanisms at play in choosing levels of sanctions. In particular, does feedback on deterrence appear to have effects on choices of levels of punishment? Does drawing people's attention to the cost of sanctions modify their choices? Does professional training change the methods of decision-making employed by individuals?

Behavioral and social scientists have increasingly gone beyond studying how aggregate outcomes come about, and have taken a plunge into the practice of using their skills to to help design "rules of the game" that achieve normatively desired outcomes.[38] We note that, especially in the case of punishment institutions, it seems that effective rules of the game will depend on the psychological motivations of the players. This is particularly stark if we consider the difference in assumptions that individuals punish for public goods motives (theories of deterrence, incapacitation) or for private benefits (cold glow). In the former case, punishment will be under provided due to free-riding motivations and so mechanisms which subsidize the costs of punishment decisions will improve overall efficiency. However, if individuals are motivated by cold glow, the same subsidies may lead to highly inefficient outcomes. Economics as "rule design" is a growing and important part of modern social science and we hope that our results contribute to this important conversation.

[38]For a survey of recent work in the field of market design see Roth (2003).

Table 2.1: Experiment 1 - Taking behavior, by condition

	(1) No vs. With Sanctions	(2) Punishment Cost Structure
1=With Sanctions	-0.655**	
	(0.0371)	
Public		-0.0556
		(0.0728)
Constant	0.841**	0.219**
	(0.0256)	(0.0625)
Observations	2407	1067

Results clustered at the subject level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 2.2: Experiment 1 - Length of punishment, by treatment, across rounds

	(1) All	(2) All	(3) Rounds 6-20	(4) Rounds 1-5
Public	1.818*	1.809*	2.113**	0.990
	(0.754)	(0.754)	(0.771)	(0.840)
Round		-0.0406*		
		(0.0186)		
Constant	1.734**	2.166**	1.394**	2.686**
	(0.455)	(0.530)	(0.457)	(0.606)
Observations	1067	1067	782	285

Results clustered at the subject level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 2.3: Experiment 1 - Length of punishment: public motives vs. cost structure

	(1) Public vs. Private	(2) Deterrence vs. None	(3) Both Effects
Public Costs	0.780*		1.818*
	(0.357)		(0.752)
No Deterrence		-1.039*	-1.039
		(0.459)	(0.767)
Constant	1.734**	3.553**	1.734**
	(0.133)	(0.148)	(0.454)
Observations	520	691	1139

Results clustered at the subject level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 2.4: Experiment 1 - Length of punishment: individual differences

	(1)	(2)	(3)	(4)	(5)
Public	1.887*			1.777*	2.317**
	(0.778)			(0.750)	(0.823)
Punishment Chosen	0.579	-1.214^{+}	1.936*		
	(0.640)	(0.686)	(0.936)		
Stolen From				-0.369	
				(0.287)	
Low Taker					-2.305**
					(0.806)
Constant	1.437*	2.358**	2.789**	1.896**	1.992**
	(0.640)	(0.733)	(0.587)	(0.515)	(0.486)
Observations	1067	448	619	1067	1067

Results clustered at the subject level

Low Taker: took less than 15 times in the no punishment condition

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 2.5: Experiment 2: Choice of punishment type, by treatment

	(1) Punish	(2) Level, Full sample	(3) Level, if Punish = 1
1 = High	-0.131	-0.154	0.584
	(0.0799)	(0.744)	(0.741)
1 = Female	-0.0296	-0.788	-0.749
	(0.0817)	(0.760)	(0.753)
Age	-0.0000964	0.0140	0.0156
	(0.00439)	(0.0409)	(0.0397)
Constant	0.938**	4.121**	4.396**
	(0.139)	(1.291)	(1.277)
Observations	81	81	69

Standard errors in parentheses

High: found with a 90% chance; Low: found with a 33% chance.

Punish=1 if assigner entered a positive level of punishment.

Level = amount of punishment chosen

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 2.6: Experiment 2: MAPP, by treatment

	(1) Take	(2) Level, Full Sample	(3) Level, if Take = 1
1 = High	0.109	-0.593	-1.799*
	(0.0991)	(0.722)	(0.813)
1 = Female	-0.211+	-1.981*	-2.109*
	(0.107)	(0.780)	(0.921)
Age	-0.00490	-0.0428	-0.0645
	(0.00573)	(0.0417)	(0.0506)
Constant	0.862**	5.322**	7.775**
	(0.179)	(1.307)	(1.616)
Observations	82	82	58

Standard errors in parentheses

High: found with a 90% chance; Low: found with a 33% chance.

MAPP = Maximum Acceptable Possible Penalties

Take=1 if taker entered a positive level of acceptable punishment.

Level = amount of acceptable punishment

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 2.7: Experiment 2: Punishment choice, all data pooled (control)

	(1) Extensive: Punish	(2) Intensive: Level
1 = High	-0.0728	-0.0692
	(0.0624)	(0.527)
1 = Assigner First	0.0290	0.939^{+}
	(0.0616)	(0.520)
1 = Female	0.0809	-0.256
	(0.0637)	(0.538)
Age	-0.00542^{+}	-0.0315
	(0.00316)	(0.0267)
Constant	0.986**	4.251**
	(0.112)	(0.944)
Observations	147	147

Standard errors in parentheses

High: found with a 90% chance; Low: found with a 33% chance

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 2.8: Experiment 2: Target's opinion on fair punishment level, by treatment

	(1) With Deterrence	(2) No Deterrence	(3) Comparing Conditions
1 = High	0.620	0.180	0.438
	(0.715)	(0.846)	(0.539)
1 = Female	-0.186	-0.154	-0.193
	(0.817)	(0.866)	(0.582)
Age	-0.0899*	-0.0536	-0.0736**
	(0.0396)	(0.0416)	(0.0282)
1 = Assigner First			-0.758
			(0.535)
Constant	6.972**	6.947**	7.374**
	(1.190)	(1.387)	(0.954)
Observations	80	64	144

Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 2.9: Experiment 3: 2nd punisher's choice, by 1st punisher choice

	(1) Full sample	(2) Crowding Out
Player 1 sanction	-0.0289	-0.569**
	(0.0620)	(0.0585)
Constant	2.199**	3.380**
	(0.237)	(0.363)
Observations	553	196

Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

2.9 Summary of Experiments

Table 2.10: Summary of experiments

Experiment	Conditions	Hypotheses tested
Experiment 1	Private costs of punishment	Role of private costs
Cost Structures	Public costs of punishment	vs. public costs
	One Round Take	No social motives
Experiment 2	Ex-ante vs. ex-post	Effects of probability of capture
Probabilities	$p = .33$ vs. $p = .9$	Ex-ante vs. ex-post behavior
Experiment 3	2 assigners	Crowding out behavior
Multiple punishers	Computer + assigner	

2.10 Summary Predictions, Experiment 1

Table 2.11: Experiment 1: Behaviors Predicted by Various Punishment Theories

Punishment Motive	Result
Incapacitation	Public Punishment = 0
	One Round Take = 0
General Deterrence	Public Punishment = 0
	Private Punishment = 0
	One Round Take =0
Specific Deterrence	Punishment decreases in rounds
	One Round Take =0
Cold Glow	Public Punishment $>$ Private Punishment
	One Round Take > 0

2.11 A Mathematical Model of Specific Deterrence

The theory of specific deterrence which we will model here is as follows: individuals start with a propensity to choose take in every round, each individual has a type $\theta \in [0, \theta^{max}]$ and a threshold level of punishment that depends on his type. The probability distribution over types is given by $p \in \Delta([0, \theta^{\max}])$ and is smooth and well behaved with density f that has strictly negative first derivative (that is, higher types are rarer).

If an individual of type θ receives a punishment of size at least θ, he 'learns his lesson' and never takes again. If he receives a punishment of size less than θ he continues to take in all rounds after.

To formalize our theory we consider a group of N honest individuals with one individual i who has been found out for taking and follows the behavioral rule outlined above, there are k rounds left in the game. We consider a benevolent social planner who does not know the type θ of the taking individual. The social planner wants to maximize the monetary rewards that will accrue to honest individuals. We now ask, given such assumptions, what can we say about the optimal punishment strategy? For simplicity, we suppose that punishments can be delivered in continuous amounts $c \in [0, \infty)$ and has a social cost of v per unit to make the math easier.

Proposition 3. *There exists a unique optimal punishment level c^* which is given by the first order condition:*

$$3f(c^*)k = v.$$

c^* is decreasing in both number of rounds left and public cost of punishment.

The intuition for the first-order condition is as follows: by marginally increasing c the social planner increases the probability that the individual in question learns their lesson from the punishment. The marginal benefit of this is exactly 3 units times the number of rounds left. The marginal cost is exactly v. When there are less rounds left, the marginal benefit is lower so optimal punishments are lower. The exact solutions, however, depend on assumptions about the distribution of types.

2.12 Proofs of Propositions

Proof of Proposition 1. The utility function P maximizes is

$$-\beta s_P + \lambda(s_P, s_T)$$

the first order conditions of the maximization are simply

$$\beta = \frac{\lambda(s *_b (\beta, s_T), s_T)}{\partial s_P}$$

which by assumption are unique (λ is concave in s_P) and give us the comparative statics directly. □

Proof of Proposition 2. Recall that we can write P's maximization problem as:

$$\psi(s_P, q)[q(\lambda(s_P, T) - \beta s_P) + (1-q)(\lambda(0, T))] + (1 - \psi(s_P, q))\lambda(0, 0).$$

We can set $\lambda(0, 0)$ to be 0 and drop the dependence of λ on the second argument to save notation. Our maximization becomes

$$\psi(s_P, q)[q(\lambda(s_P) - \beta s_P) + (1-q)(\lambda(0))].$$

Note that if we take the derivative we get

$$\psi'(s_P, q)[q(\lambda(s_P) - \beta s_P) - (1-q)(\lambda(0)] + \psi(s_P, q)[q(\frac{\partial \lambda}{\partial s_P} - \beta)].$$

The first term is positive because ψ' is negative and the quantity in parentheses which it multiplies is negative from the assumption that $\delta(s_P) > 0$.

Now, consider the ex-post problem with the same λ. The answer to this problem is given $\overline{s}$ that sets

$$\beta = \frac{\partial \lambda}{\partial s_P}$$

this means that for $s < \overline{s}$ we have that the second term must be also positive and hence the overall utility only increases for $s \in [0, \overline{s}]$ so any maximizer of the ex-ante problem must be above the maximizer of the ex-post problem. Additionally, we may have that the original maximization problem has several local maxima (and hence we cannot, without more

conditions, describe the maximum using derivatives), however it is a continuous function on a convex set so it will generically have one global maximum which is, by the argument above, guaranteed to lie about $\overline{s}$.

□

Chapter 3

Ambiguity, Information and Valuation

3.1 Brief Summary

We use ambiguous financial gambles to explore how information affects individual decisions. In the domain of gains, we find that information that suggests that a gain is more likely increases individual willingness to pay (WTP) for the gamble as well as their confidence. In contrast, information that suggests a gain is less likely has a much smaller effect on both WTP and confidence. We find a similar imbalance in the domain of losses: information which suggests a loss is less likely influences behavior much more than information which suggests a loss is more likely. We show that two mechanisms appear to drive this effect: a bias towards integration of favorable information as well as an "unshrouding" effect in which unfavorable information has a component which has a positive effect on WTP: removing a part of ambiguity from the decision problem.

3.2 Introduction

Most decisions in the real world, whether they are the decisions made by a trader to buy a stock or an individual deciding whether to visit a particular restaurant have outcomes which are only probabilistically known. Decisions in this domain have been characterized using two dimensions: risk, or the probability of a given outcome, and ambiguity, the availability or unavailability of necessary information to make these risk estimates. Beyond their

response to risky situations, individuals tend to be ambiguity averse (Ellsberg (1961), Fox and Tversky (1995)). Much is known about how individuals respond to changes in risk levels (Roth and Kagel (1995), Kahneman and Tversky (2000)). However much less is known about how addition of information affects ambiguous decisions.[1] Given the prevalence of ambiguity, answering this question is a major component of understanding and predicting important human behaviors (Knight (1921), Epstein and Schneider (2008), Ritov and Baron (1990)). Here we report a series of experiments studying the effects of information on ambiguous decisions using financial gambles.

Traditional experiments on ambiguity aversion are modeled on what has come to be known as the Ellsberg paradox (Ellsberg (1961)). In these experiments, participants are presented with a bag containing 100 poker chips, all of which are either red or blue. They are then asked for their willingness to pay (WTP) to play a game in which they guess the color of a chip drawn from the bag at random. Players win a monetary reward if the color of the drawn chip matches their guess, but receive nothing if it doesn't. On average individuals are willing to pay more when they know the bag contains exactly 50 red and 50 blue poker chips (no ambiguity) than when they know nothing about the bag's contents (complete ambiguity). This occurs even though they maintain the same risk estimate for a chip of their color being drawn which contradicts the predictions of subjective expected utility models.

We extended this design to investigate how information influences the evaluation of ambiguous situations. In a series of experiments participants were asked to indicate their WTP for tickets to play several iterations of one of two types of games. In "pull-a-chip" games a poker chip was randomly drawn from a bag containing 100 red and blue colored chips. A red chip resulted in a winning outcome with a (hypothetical) monetary payout. In "majority" games a bag was filled with 101 poker chips and participants won if 51 or more of the chips in the bag were red. Note that in the majority game, but not in the pull-a-chip game, perfect knowledge of a bag's contents guarantees knowledge of whether

[1]There is much experimental work on behavioral foundations of ambiguity aversion (Halevy (2007) is a recent example and Camerer and Weber (1992) provides a survey of the older literature) but, to the best of our knowledge, none that considers anything similar to the information variation done in the experiments reported here.

the outcome would be a win or loss. Thus, the majority game involves only ambiguity whereas the pull-a-chip game involves ambiguity together with risk. In our experiments, in addition to reporting their WTP, participants were asked to rate how confident they were that a winning chip would be drawn, or that the winning color would be the majority color on 7-point Likert scales.

For each bag, participants were given partial information about its contents in the form of the following statement: "This bag contains at least X red chips and at least Y blue chips." X and Y were varied parametrically from 0 to 50 in increments of 25, thus creating 9 possible levels of knowledge, or rounds, for each game. Across both hypothetical and incentivized experiments we found that favorable information (increases in number of known winning chips) increased WTP and confidence while unfavorable information (increases in number of known non-winning chips) had a significantly smaller effect on WTP and confidence.[2]

Additionally, in order to see whether our effect persisted in contexts where information could be interpreted more subjectively, participants played iterations of another game which was based on evaluating trivia. In these games, participants indicated their WTP for gambles dependent on trivia about categories such as sports, geography, finance, weather and knowledge of cities. Participants won a monetary payout if the trivia item in question was true and received nothing if it was false. Participants were also randomly assigned 0, 1 or 2 facts pertaining to the trivia and made their WTP decisions in the presence of all this information. After indicating their WTP, participants also evaluated the facts assigned to them, giving a subjective assessment of whether each fact was favorable ("this fact makes me think the trivia is more likely to be true"), unfavorable ("this fact makes me think the trivia is less likely to be true"), or neutral ("this fact has no effect on my estimate"). Just as in our poker chip games, favorable information increased WTP while unfavorable information had a much smaller effect.

While this series of experiments demonstrate a strong and consistent bias in ambiguous situations involving gains, it is less clear how such a bias might translate to ambiguous

[2]Magnitude levels of unfavorable effects relative to favorable effects varied between 0% and 30% depending on experimental population.

situations where the outcome reflects avoiding or minimizing a loss. To address this, we conducted an additional (hypothetical) experiment in the domain of losses. Participants started with a specific endowment and a bag with 100 poker chips. A chip would be drawn at random and their endowment would be lost if a red chip were drawn, but retained if a blue chip were drawn. Participants were asked to enter their WTP for an insurance ticket which would protect their endowment if a red chip were drawn. We found that an increase in number of no-loss chips significantly decreased WTP for insurance, but knowledge about the number of loss chips had no significant effect.

Our final series of analyses look at what psychological mechanisms drive this asymmetry. First, we recruited individuals to simply rate the trivia statements from experiment 2 as being favorable or unfavorable without the addition of any sort of gamble incentive. This let us classify trivia facts as 'objectively favorable' or 'objectively unfavorable.' We found that participants in the trivia gambling experiments agree with the third party ratings 75% of the time when the ratings are objectively favorable but only 50% of the time when the ratings are objectively unfavorable. This provides a hint of biased integration.

The difference between completely ambiguous and risk only situations has sometimes been discussed to be caused by relative ignorance (Fox and Tversky (1995)). That is, ambiguous gambles are aversive (and thus are valued less than subjectively equivalently risky ones) because they involve a negative affective component caused by lack of knowledge. In the language of a Bayesian model, this can be framed as individuals not simply caring about the 'point estimate' of their prior but also about something like its entropy.[3] Because of this, there is another channel by which our asymmetric effect may arise: each piece of information has two effects, it changes point estimates and it decreases entropy. For favorable information, both of these effects go in the same direction (both positive) while for unfavorable information one effect is negative (decrease in point estimate) but the other is positive (decrease in entropy).

To see if such an effect existed, we ran a final experiment in which individuals were presented with our pull-a-chip gambles but were now asked to indicate three things: their

[3]Indeed recent neuroscientific (Hsu et al. (2005), Huettel et al. (2006)) studies seem to bear out the idea that ambiguous and purely risky situations are fundamentally different.

WTP, their 'subjective estimate' that a red or blue chip would be drawn (i.e. their point estimate) and their subjective feeling of how 'accurate' their subjective estimate is (i.e. the entropy of their prior). We show that individual WTP can be broken down into a combination of point estimate minus entropy. Additionally we show that the effects of information are of expected directions: favorable information increases likelihood estimates and increases subjective accuracy while unfavorable information decreases likelihood estimates but also increases subjective accuracy. We find some evidence of a bias in integration towards favorable information as well.

We now turn to describing our experiments and results. First we describe experiment 1 and show our main behavioral effect. We then discuss three controls. We then replicate our effects in an incentive compatible environment (experiment 2, part 1) as well as in our trivia paradigm (experiment 2, part 2). Finally, we show that our effect also holds in a loss frame. We then discuss possible mechanisms driving our effect and conclude with a brief sketch of theoretical implications.

3.3 General Methods

As many of our experiments use the same stimuli, we discuss the construction of them here for simplicity.

To produce 'rounds' for each of our experiments, we combined a game type (pull-a-chip: 100 poker chips, red or blue, winning condition = one chip drawn at random matches your assigned color; majority: 101 poker chips, red or blue, winning condition = majority color of bag matches your assigned color) with a knowledge level or a statement of the form "This bag contains at least X red chips and at least Y blue chips." X and Y were varied parametrically from 0 to 50 in increments of 25 to create 9 different knowledge levels (table 3.1).

Some of our experiments use participants recruited from Amazon's Mechanical Turk (AMT) while others take place at the Harvard Decision Science Lab, we discuss this in the methods section of each experiment. In experiments which used AMT, the knowledge levels were given in the sentence above. In lab experiments, individuals saw both the verbal

description as well as a pictorial representation (see online Appendix for sample stimuli). We note that our experiments differ from some traditional experiments in that winning colors are assigned and not chosen by participants – this is done so that we can observe behavior when unfavorable information is in the majority. A control experiment which allows choice shows our results are not artifacts of this particular design change.

Table 3.1: Knowledge levels used in experiments

No information	At least 25 red	At least 50 red
At least 25 blue	At least 25 red At least 25 blue	At least 50 red At least 25 blue
At least 50 blue	At least 25 red At least 50 blue	At least 50 red At least 50 blue

In all experiments, after providing informed consent in accordance with the IRB standards of the supporting university, participants indicated basic demographic information (age, gender). In addition, participants recruited from AMT completed attention checks that including copying text into a response box as well as answering basic questions about the rules of the experiment (eg. "In the game, how many poker chips are in each of the bags?")

3.4 Experiment 1: Main Study

3.4.1 Methods

Two hundred and fifty-six participants (48 $\%$ male, mean age = 31.48) recruited via AMT participated in this study for monetary compensation. Participants were paid 30 cents for "playing hypothetical lottery games." Out of the full set, forty-five participants (17 $\%$) were unable to successfully complete one of the attention checks, and were removed from the analysis, leaving a sample size of n = 215.

This experiment used a joint evaluation design and prior testing revealed that including too many rounds on a single page was confusing. In this experiment each participant viewed 5 bags simultaneously (see sample stimuli section for example). The experimental

design was constructed as follows: participants were randomly assigned into one of six conditions ($\sim$ 37 participants per condition) in a between-subject design by combining a game type (pull-a-chip vs. majority) with a subset of these bags. In majority games, participants were asked to imagine five bags (labeled as Bags A through E), each filled with 101 poker chips colored red or blue. In pull-a-chip games, participants were asked to imagine five bags (labeled Bags A through E), each filled with 100 poker chips colored red or blue. Participants were given varying amounts of information about the specific composition of the chips in each bag. Table 3.2 shows the possible bag sets. Note that these bag sets combined with game types include all 18 possible levels of knowledge.

Table 3.2: Possible bag sets

	Set 1	**Set 1**	**Set 3**
Bag A contains $\geq$...	50 red, 50 blue	50 red, 50 blue	50 red, 50 blue
Bag B contains $\geq$...	25 red, 25 blue	25 red, 25 blue	25 red, 25 blue
Bag C contains $\geq$...	25 red	50 red	25 red, 50 blue
Bag D contains $\geq$...	25 blue	50 blue	50 red, 25 blue
Bag E contains $\geq$...	unknown	unknown	unknown

Participants were asked to enter their hypothetical WTP for a ticket to play pull-a-chip games or majority games in each condition for a prize of $\$50$. Participants indicated this WTP for each of the five bags by moving sliders on a scale that ranged from $\$0$ to $\$35$. Afterwards on a separate screen, they reported how confident they felt that a red chip would be drawn on a scale of 1(Not Very Confident) to 7(Very Confident).[4]

3.4.2 Results

This design contained a test of the original Ellsberg paradox. Consistent with existing results average WTP was lower for a pull-a-chip bag with no color composition information given (mean = 6.31) compared to a pull-a-chip bag with a known composition of 50 red chips and 50 blue chips (mean = 13.37). This difference was highly significant (t-stat = -6.02, $p < .001$).

[4]The question used, for each bag, was: "How confident do you feel that a red chip will be drawn (red will be the majority color) in this bag?"

Figure 3.1 shows WTP pooled across games for each possible knowledge level. At first glance, the impacts of favorable and unfavorable information seems to be highly different: favorable information seems to increase WTP while unfavorable information seems to have a much smaller effect.

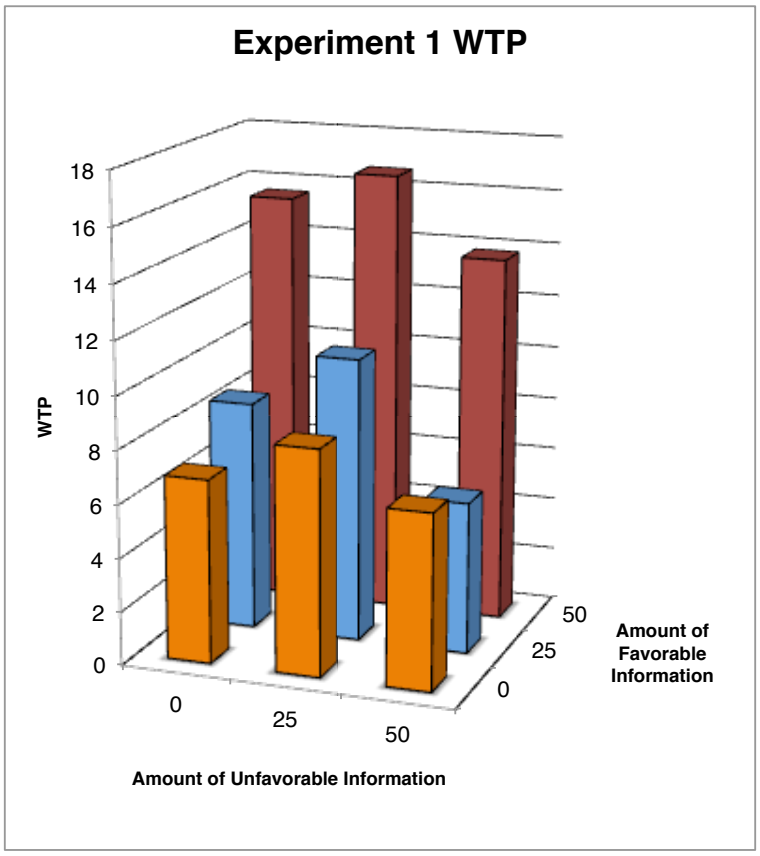

Figure 3.1: Experiment 1 WTP by knowledge levels

To analyze our data, we used linear regression. In all reported regressions standard errors are clustered at the participant level to take account for intraindividual correlations in WTP.

To form our base specification, we pooled data from all games and knowledge levels and regressed willingness to pay (WTP) on numred (number of red chips, the winning color), numblue (the number of blue chips, the non-winning color). Column 1 of table 3.3 shows our main results, both the effects of favorable and unfavorable information are significant

however the magnitude of the effect of unfavorable information is approximately 13% of the effect of favorable information. A Wald test rejects equality of the absolute values of the two coefficients at a level of $p < .01$ – thus favorable information has a much larger effect on WTP than unfavorable information.

In robustness checks we included a dummy variable for the game type (column 3), as well as interactions between the game type and numred and numblue. In our third specification, we added age and gender as regressors as well as gender interacted with numred and numblue (column 3). We control for these variables as there is existing evidence that men and women differ in their risk preferences (Borghans et al. (2009)) and that risk preferences also differ by age (Tymula et al. (2012)). Finally we used a Tobit model (column 4) to make sure that our results were robust to censoring. Our main effect is robust across all specifications. One additional finding of note is that men appear to respond more favorably than women to positive information when evaluating WTP. This is consistent with evidence that men tend to be less risk averse than women, however, no further interpretation of this result is possible given this experimental design.

Tables 3.15 and 3.14 in the Appendix show the same regressions independently performed on majority games and pull-a-chip games respectively. These regressions replicate our primary finding and are robust to including subject fixed effects (column 4). This analysis also indicates that the negative effect of unfavorable information is driven specifically by majority games (a finding which is replicated in our later lab experiment).

These regressions used continuous variables to represent the number of chips. This facilitates interpretation of regression coefficients. Specifically the coefficient on favorable is the average marginal effect of an extra piece of favorable information, and similarly for the coefficient on unfavorable.

We also replicated the analysis using a factor model (Table 3.4). We regressed WTP on dummy variables for numred = 25, numred = 50, numblue = 25, numblue = 50 (column 1). As robustness checks we included interactions of these dummies (column 2), the addition of game-type (column 3), gender and age (column 4) and a Tobit to check for censoring (column 5).

Our main claim, that favorable and unfavorable information have asymmetric effects, is robust to this analysis. However, we note effects of favorable information seem to be

Table 3.3: Experiment 1: WTP on continuous variables

	(1) wtp	(2) wtp	(3) wtp	(4) Tobit
numred	0.161**	0.169**	0.122**	0.143**
	(0.0141)	(0.0207)	(0.0237)	(0.0257)
numblue	-0.0210*	-0.0390**	-0.0257+	-0.0310+
	(0.00892)	(0.0129)	(0.0144)	(0.0167)
pull		-1.279	-1.385	-1.283
		(0.971)	(0.971)	(1.097)
pullXred		-0.0118	-0.00310	0.00166
		(0.0279)	(0.0277)	(0.0305)
pullXblue		0.0373*	0.0327+	0.0319
		(0.0178)	(0.0176)	(0.0203)
gender		3.957**	2.481*	2.993**
		(0.960)	(0.986)	(1.115)
age		-0.0613	-0.0613	-0.0690
		(0.0457)	(0.0458)	(0.0494)
genderXred			0.0768**	0.0714*
			(0.0270)	(0.0296)
genderXblue			-0.0195	-0.0231
			(0.0173)	(0.0200)
Constant	6.703**	7.044**	7.929**	7.064**
	(0.496)	(1.659)	(1.643)	(1.765)
Observations	1075	1075	1075	1075

Standard errors in parentheses clustered at participant level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

convex as the coefficient on favorable = 50 is more than twice the size of the coefficient on favorable = 25. We also note that the effect of unfavorable information seems to be non-linear as the addition of a small amount of unfavorable information (unfavorable = 25) seems to increase WTP while the addition of a large amount (unfavorable = 50) seems to decrease WTP, though again, this effect is much smaller in magnitude than positive effect of favorable.

Given that in our main discussion is about average treatment effects (that is, average marginal effects of favorable/unfavorable info), we report continuous regressions for our three controls from here forward. We will return to the topic of convexity when discussing the results of experiments 2 and 3.

We performed the same regression analyses for confidence ratings and again find that favorable information has a much stronger effect than unfavorable information (see table 3.16 in appendix). As there is no main effect of game or interaction effects, we do not report regressions on confidence separated by game type. In addition, we find no qualitative difference in the effects of information between different bag sets.

3.5 Experiment 1: Choice Control

We also note that our design differs from several prior experiments on ambiguity aversion in which the winning color is pre-assigned. We performed an additional control study in which participants still indicated their WTP, but could choose the winning color of their ticket.

3.5.1 Methods

One hundred and sixty-six individuals (66% male, mean age = 28.47 s.d. = .748) recruited via AMT participated in this study for monetary compensation. Eight participants were unable to complete our attention check and were removed from the analysis leaving a sample size of n=158.

Participants were assigned into one of two conditions, pull-a-chip (n=79) or majority (n=79) and presented with hypothetical bags of poker chips as well as partial informa-

Table 3.4: Experiment 1 WTP using factor model

	(1) wtp	(2) wtp	(3) wtp	(4) Tobit
numred=25	1.359**	1.773*	1.845**	2.177**
	(0.521)	(0.714)	(0.699)	(0.779)
numred=50	8.228**	8.677**	8.539**	9.602**
	(0.708)	(1.241)	(1.195)	(1.334)
numblue=25	1.752**	1.583	1.295	1.228
	(0.490)	(1.061)	(1.056)	(1.261)
numblue=50	-1.566**	-0.350	-0.488	-0.926
	(0.447)	(0.851)	(0.862)	(1.056)
numred=25 X numblue=25		0.387	0.392	0.648
		(1.293)	(1.262)	(1.442)
numred=25 X numblue=50		-2.646*	-2.153+	-2.394
		(1.219)	(1.236)	(1.491)
numred=50 X numblue=25		-0.497	0.355	0.454
		(2.147)	(2.079)	(2.318)
numred=50 X numblue=50		-1.405	-1.130	-0.935
		(1.523)	(1.452)	(1.610)
pull		-1.072	-0.725	-0.626
		(0.969)	(0.930)	(1.038)
gender			3.830**	4.271**
			(0.967)	(1.070)
age			-0.0630	-0.0743
			(0.0456)	(0.0500)
Constant	7.087**	7.398**	7.134**	6.487**
	(0.479)	(0.732)	(1.707)	(1.847)
Observations	1075	1075	1075	1075

Standard errors in parentheses clustered at participant level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

tion about their contents similar to the conditions in experiment 1 in a joint evaluation paradigm.. We only chose a subset of conditions for this control: $(R = 50, B = 50), (R = 50, B = 0), (R = 25, B = 25), (R = 0, B = 0)$ for pull-a-chip games and $(R = 50, B = 50), (R = 50, B = 0), (R = 25, B = 25), (R = 25, B = 0), (R = 0, B = 0)$ for majority games.[5]

For each bag, participants chose the color that defined a winning outcome and entered their WTP. Participants also indicated their confidence in winning the game on 7 point Likert scale.

3.5.2 Results

Based on the participants' choice of winning color we regressed their WTP for a ticket to that game on the number of chips of the same color (favorable) and number of chips of the non-winning color (unfavorable). We used the same regression specifications as in experiment 1.

Table 3.5 shows the results using willingness to pay as the dependent variable, and table 3.17 shows the same results using confidence ratings as the dependent variable. Our analysis shows that favorable information has a significant effect on both WTP and confidence, while the effect of unfavorable information has a negligible effect. This replicates the findings in experiment 1, and suggests that our effects are not influenced by whether individuals can choose the winning color themselves. Table 3.17 in the appendix shows the same regressions for confidence levels, again, replicating the main study effects.

We also note that participants seem to use the information to make their choices: when participants have symmetric knowledge (that is, they know that the bag contains at least X red and at least Y blue chips and $X = Y$) they choose red as their winning color 53% of the time (not statistically different from 50%). When $X > Y$ they choose red 66% of the time (statistically different from 50%) and when $X = 50$ and $Y = 0$ they choose red 86% of the time (again, statistically different from 50%).

[5]The reason for these choices is actually a chronological one, we ran the control first, found results with choice and decided to expand that to what later became experiment 1 with the assigned colors.

Table 3.5: Experiment 1 (with choice): WTP on continuous variables

	(1) wtp	(2) wtp	(3) wtp	(4) wtp
chosencolor	0.114**	0.0901**	0.0684*	0.0779*
	(0.0141)	(0.0234)	(0.0293)	(0.0317)
othercolor	0.00314	-0.0149	-0.0178	-0.0119
	(0.0126)	(0.0177)	(0.0199)	(0.0201)
pull		0.447	0.372	0.214
		(1.046)	(1.050)	(1.154)
pullXchosen		0.0384	0.0412	0.0465
		(0.0285)	(0.0285)	(0.0301)
pullXother		0.0499*	0.0507*	0.0537*
		(0.0241)	(0.0242)	(0.0246)
gender		4.451**	3.618**	3.910**
		(1.017)	(0.992)	(1.139)
age		-0.0455	-0.0461	-0.0435
		(0.0613)	(0.0613)	(0.0640)
genderXchosen			0.0301	0.0288
			(0.0278)	(0.0299)
genderXother			0.00350	0.000757
			(0.0215)	(0.0215)
Constant	6.935**	5.095*	5.705**	4.753*
	(0.556)	(2.166)	(2.102)	(2.275)
Observations	711	711	711	711

Standard errors in parentheses clustered at participant level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

3.6 Experiment 2: Incentive Compatible

Up to this point, our results have reflected responses to hypothetical gambles. Since decisions with real consequences are more representative of daily decision-making, and can elicit stronger reactions than hypothetical scenarios we conducted an incentive compatible experiment. Thirty-seven undergraduate students participated in a lab-based study using the same set of games and possible knowledge levels as in experiment 1 in a within-subject design using real monetary outcomes. The experiment consisted of two semi-independent parts, each subdivided into rounds. To keep exposition flowing, we will first discuss the methods and results of part 1, then move to part 2.

3.6.1 Poker Chips (Incentive Compatible Control)

In each round of part 1, participants entered their WTP for pull-a-chip or majority games from bags of poker chips about which they received partial information. Just as in experiment 1, winning colors were assigned, however in this experiment the winning color (red or blue) was randomly assigned for each round and participants were informed of this.

Each participant made a total of 18 choices (9 levels of knowledge x 2 types of games) in part 1. The rounds were presented in random order with the following exception: the first evaluation for all participants was the pull-a-chip bag with no information and their second evaluation was the pull-a-chip bag whose contents were known to be 50 red chips and 50 blue chips. Additionally, participants were asked for their confidence that they would win each game using the same 1 to 7 scale as experiment 1. The Appendix includes instructions and sample stimuli.

A version of the Becker-DeGroot-Marschak (BDM) procedure was used to aid in elicitation of true WTP (Becker et al. (1964)). Participants were given an endowment at the start of each round and had their WTP compared with a price randomly generated by the computer. If the computer's price was higher than their stated WTP, they retained the full endowment for that round. If the computers price was lower, they bought the ticket out of their endowment for that price. The BDM was (painstakingly) explained to participants along with a demonstration of the optimal strategy. We note that since most of our study's

power comes from the 18 within-subject questions and what is at stake are changes rather than levels common misunderstandings of the BDM (for example, subjects treating it as a first price auction and consequently shading all bids by a percentage) will not bias our results.

To prevent "portfolio building" only one round of the experiment (that is, from combined rounds in both part 1 and part 2) was chosen to count for real money and was actually played out at the end of the experiment. This involved conducting the BDM procedure and determining or revealing the outcome of the round. No deception was used and all 18 actual bags of poker chips described in the experiment were shown to subjects before they began.

Our data is again consistent with standard tests of ambiguity aversion: WTP for a pull-a-chip bag with no information (M = 6.27 s.d. = .88) was significantly lower than that for a pull-a-chip bag with a known composition of 50 red and 50 blue chip (M = 10.40 sd = .86; t-test $p < .001$).

In our base regression (table 3.6) we regressed WTP on numgood (number of favorable known chips) and (number of unfavorable known chips) and replicated the qualitative findings of experiment 1. The absolute magnitude of the coefficient on unfavorable information was approximately 30% that of the coefficient on favorable information. A Wald test rejects equality of these coefficients at $p < .01$.

A second regression included a variable indicating the type of game as well as interactions of numgood and numbad with this dummy (column 2). This specification reveals that the significant negative effect of conflicting information is primarily driven by majority games, as the coefficients on the interaction term pullXbad is significant and of the same magnitude as the negative coefficient on numbad.[6] We also include gender, age and gender interacted with information types (column 3) as well as subject level fixed effects (column 4). Finally, we use to a Tobit regression to show our effects are not driven by censoring (column 5). We note that we find no gender effects here unlike in experiment 1. We also did the same analysis for confidence ratings (table 3.20 in appendix) which replicate the results of experiment 1.

[6]We can also disaggregate the WTP behavior by game type. Table 3.19 and 3.18 in the Appendix show the same regressions for pull-a-chip and majority games separately. All of these effects can be seen in the separate regressions as well.

Table 3.6: Experiment 2 (Poker Chips): WTP on continuous variables

	(1) wtp	(2) wtp	(3) wtp	(4) (with FE)	(5) Tobit
numgood	0.117**	0.136**	0.118**	0.135**	0.137**
	(0.0123)	(0.0147)	(0.0169)	(0.0152)	(0.0189)
numbad	-0.0352**	-0.0581**	-0.0471**	-0.0585**	-0.0542**
	(0.00998)	(0.0143)	(0.0141)	(0.0147)	(0.0174)
pull		-0.142	-0.142	-0.153	0.132
		(0.529)	(0.532)	(0.545)	(0.650)
pullXgood		-0.0363*	-0.0363*	-0.0360*	-0.0485**
		(0.0142)	(0.0143)	(0.0146)	(0.0169)
pullXbad		0.0459**	0.0459**	0.0462**	0.0572**
		(0.0136)	(0.0137)	(0.0140)	(0.0171)
ismale			-1.170		-1.555
			(1.439)		(1.770)
age			0.172		0.209
			(0.214)		(0.235)
maleXgood			0.0451+		0.0599+
			(0.0251)		(0.0310)
maleXbad			-0.0292		-0.0401
			(0.0219)		(0.0265)
Constant	6.432**	6.504**	3.205	6.517**	1.690
	(0.640)	(0.739)	(4.586)	(0.455)	(5.032)
sigma					
Constant					5.948**
					(0.274)
Observations	665	665	665	665	665

Standard errors in parentheses clustered at participant level.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

We can also consider a factor analysis as in experiment 1. We regress WTP on dummies for favorable=25, favorable=50, unfavorable=25, unfavorable = 50 as well as other control variables (table 3.7). We now find much stronger convexity in our results than in experiment 1: the addition of a small amount of favorable information (25) seems to have either no effect or a very small positive effect, while the addition of a lot of favorable information (50) seems to have a much larger effect. Again, unfavorable information seems to be non-monotonic as a small amount of unfavorable information seems to *increase* WTP while a larger amount seems to decrease it (relative to having no unfavorable information). We do not speculate on the causes of this,[7] we note that while this is not a problem for our main argument which is that favorable and unfavorable information affect decisions quite differently. Favorable information seems to increase WTP (though perhaps convexly) while unfavorable information seems to do all sorts of weird things. Even dropping the intermediate conditions (favorable=25, unfavorable=25) so that we restrict to conditions where both effects are in the expected direction, the effect of favorable information seems to be much stronger than the countervailing force of unfavorable information.

3.6.2 Trivia Questions

In part 2, individuals saw 10 pieces of trivia in a random order. For each trivia item, they were randomly assigned to see either 0,1 or 2 related facts (see sample stimuli for example, the Appendix lists both the trivia items and possible associated facts). Participants were informed of the following rules: if the trivia statement was true, they would win additional money, if it was false, they would win nothing. After indicating WTP, participants were asked to indicate their confidence that they would win on a scale of 1 to 7 as well as how knowledgeable they felt about the topic of the trivia question on a scale of 1 (Not very knowledgeable) to 7 (Extremely knowledgeable).

Each trivia question was in the form of a 'less than' or 'more than' statement and we

[7] In future research we plan to explore the non-monotonicities in unfavorable information, however our current experimental designs are simply not set up to make any conclusive statements about possible causes of this non-monotonicity (and why convexity seems to be different between hypothetical and incentive-compatible designs) or even to look at it more in depth. Any explanations given here would really be post-hoc rationalizations of our data, not exactly appropriate for a study which looks at a form of confirmation bias.

Table 3.7: Experiment 2 (Poker Chips): WTP with factor model

	(1) wtp	(2) wtp	(3) (with FE)	(4) Tobit
numgood=25	-0.0270	0.757	0.757	1.337
	(0.519)	(0.631)	(0.648)	(0.843)
numgood=50	5.874**	7.284**	7.284**	8.262**
	(0.614)	(0.954)	(0.981)	(1.161)
numbad=25	0.779+	1.757*	1.757*	2.319**
	(0.393)	(0.736)	(0.757)	(0.888)
numbad=50	-1.757**	-0.541	-0.541	-0.674
	(0.500)	(1.010)	(1.038)	(1.270)
numgood=25 X numbad=25		-1.041	-1.041	-1.552
		(0.899)	(0.924)	(1.072)
numgood=25 X numbad=50		-1.311	-1.311	-1.451
		(1.012)	(1.040)	(1.280)
numgood=50 X numbad=25		-1.892+	-1.892+	-2.455*
		(1.053)	(1.083)	(1.220)
numgood=50 X numbad=50		-2.342+	-2.369+	-2.426+
		(1.225)	(1.255)	(1.467)
pull		0.0970	0.103	0.300
		(0.260)	(0.266)	(0.305)
ismale		-0.582		
		(1.085)		
Constant	6.866**	6.307**	6.086**	5.094**
	(0.645)	(0.940)	(0.632)	(1.092)
Observations	665	665	665	665

Standard errors in parentheses clustered at participant level.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

randomized the 'direction' ('less than' or 'more than') of each trivia question at the group level and participants were informed of this fact in the instructions. This was done so that participants did not feel that they were being asked 'trick questions.'

When participants received additional facts along with their trivia, they were also asked to categorize those facts as favorable ("This fact makes me think that it is more likely that the trivia is true"), unfavorable ("This fact makes me think that is it less likely that the trivia is true") or neutral ("This fact does not help me make any judgment on the trivia"). Approximately 65 $\%$ of facts are rated as non-neutral. Figure 3.2 shows the average WTP collapsed across questions by subjectively reported knowledge level.

Regression analysis for the trivia pooled all 10 questions (table 3.8). In our base specification, we regressed WTP on a dummy variable (subjgood) that takes value 1 if the participant indicated the presence of favorable information and a second dummy variable that takes value 1 if the participant indicated the presence of unfavorable information (subjbad). Both have significant and robust effects in the correct directions.

However, the relative magnitudes of the point estimates of unfavorable information is at best approximately 30% that of the coefficient on favorable information and their absolute magnitudes are significantly different ($p < .01$, Wald test). This is a strong test of biased integration as the information valence is identified by the subjects themselves. Thus even when subjects say that a fact makes the trivia item less likely to be true, they do not appear to significantly change their valuations to reflect this assessment.

Robustness checks included controls for self reported knowledgeability of the topic, gender and gender interacted with types of information (column 2) as well a specification that also included fixed effects for each trivia question (column 3). Finally we used a Tobit regression (column 4) to see that our results were not driven by censoring. The same regressions for confidence levels can be found in the appendix (table 3.21).

3.7 Experiment 3: Losses

While this series of experiments demonstrate a strong and consistent bias in ambiguous situations involving gains, it is less clear how such a bias might translate to ambiguous situ-

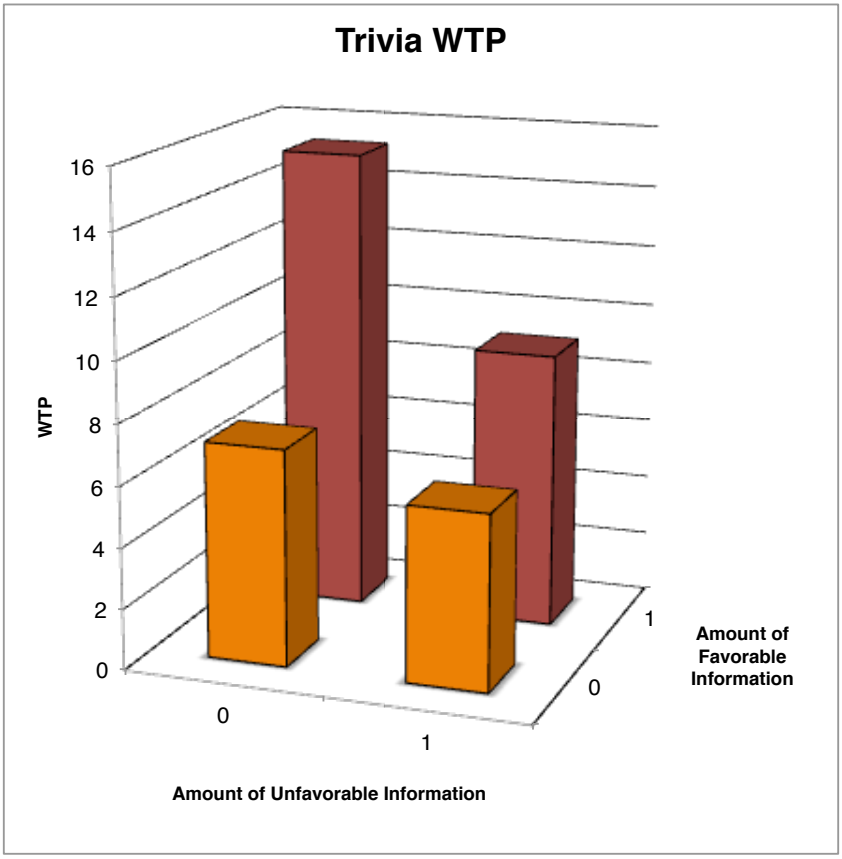

Figure 3.2: Experiment 2 (Trivia) WTP by knowledge levels

ations where the outcome reflects avoiding or minimizing a loss. We have been discussing this effect in terms of favorable or unfavorable information, however a behaviorally equivalent hypothesis could be that individuals are simple much worse at incorporating conflicting information about a salient hypothesis than confirming information.[8]

Thus if participants are simply evaluating a salient hypothesis, we conjecture that in a

[8]There is much psychological research on 'unmotivated' confirmation bias, for example in the famous Wason card task individuals are presented with 4 double sided cards. One side of each card has a color, the other side has number. Two cards are presented color up (red and brown) and the other two are presented number up (1 and 4). Participants are asked to turn over the minimal number of cards to test the hypothesis: "All cards with an even number on one face, have red on the other." Most participants correctly turn over the 4 but do not turn over the brown card (usually, they check the red card, however finding an odd number on the back of a red card *does not* tell us anything about the veracity of the original hypothesis). This task is a specific example of a general principle which could arguably be at play in our experiments: people are simply not good at modus tollens.

loss frame information that argues that a loss is more likely should have a much larger effect than information which contradicts this hypothesis. On the other hand, if what is causing the imbalance is the 'favorable' nature of the information then the effects of information which argues that a loss is less likely should dominate the effects of 'loss more likely' information. To address this, we conducted an additional experiment in the domain of losses.

3.7.1 Methods

Sixty participants were recruited from AMT and were presented with hypothetical pull-a-chip rounds in a within-subject design. Participants were presented with 9 rounds. Participants began each round with an endowment of $\$50$, at the end of a round a chip was drawn from a bag of 100 red or blue chips. If the drawn chip were red, participants would lose their endowment, if it were blue, they would keep their endowment. Participants were then asked to indicate their WTP for an insurance ticket, which would protect their endowment if a red chip were drawn. An insurance ticket had no effect if a blue chip was drawn. Additionally, participants were given partial information about each bag's contents using the knowledge levels from experiment 1. All participants entered responses for all 9 knowledge levels.

3.7.2 Results

We used continuous linear regression to measure the impact of information on WTP for an insurance ticket. Regressing WTP on number of red chips and number of blue chips shows that number of blue (no loss chips) decreases WTP for insurance while number of red (loss chips) has no significant effect on WTP (table 3.9). Using just the point estimates, average magnitude of the effect of unfavorable (loss) information is approximately 30% of the effect of favorable information with a Wald test rejecting equality of the absolute value of the coefficients at the $p < .01$ level. This result is robust to adding controls for gender, age and interactions (column 2), subject level fixed effects (column 3) as well as using a Tobit regression to check for censoring (column 4).

We again find convexity in our results. Regressing WTP for insurance on a factor model (table 3.10) with controls and interactions shows that small amounts (25 chips) of favorable information seem to have a small negative effect on WTP for insurance while large amounts (50 chips) have a larger (more than 2x the size) negative effect indicating convexity. The effects of unfavorable information seem to be all over the place. In some specifications small amounts (25 chips) of unfavorable information seems to increase WTP for insurance tickets, in some it seems to decrease it. In the most 'pessimistic' (relative to our hypothesis) specification, unfavorable information's positive effect on WTP for insurance is much smaller than the countervailing effect of favorable information. Again, this is consistent with our main argument: favorable information affect WTP significantly and in the right direction, while unfavorable information is treated qualitatively differently.

3.8 Biased Integration

In this section we begin to expand beyond simply using WTP data to look for correlations and explore psychological causes of our effect. Informally, we have been arguing that individuals underweight unfavorable information during information processing. Though our main experiments are not set up to look at the mechanism underlying our effects, here we present a very simple model and some suggestive data looking at this hypothesis.

Consider the following simple model: there is a state of the world $\theta \in \{\theta_T, \theta_F\}$. For us, this is whether the trivia statement in question is true or false. Individuals start with a prior $p(\theta_T)$ and receive a signal s (a fact presented to them) of accuracy q (that is, $prob(s \mid \theta = \theta_T) = q$). A simple reduced form way of modeling bias towards favorable information is that when individuals receive a signal of accuracy $q < \frac{1}{2}$ (that is, F is more likely) they treat it as having accuracy $\alpha q + (1 - \alpha)\frac{1}{2}$ with $\alpha \in [0, 1]$ thus reducing its impact on their posteriors when T matters for their payoffs, however, when they are simply asked to estimate probabilities, they incorporate information with no bias.

Now, suppose that individuals report whether a fact is favorable or unfavorable using a cutoff rule. That is, if $q \in [0, \frac{1}{2} - \beta]$ they report the fact as unfavorable, $q \in [\frac{1}{2} + \beta, 1]$ they report as favorable (with $\beta \in [0, \frac{1}{2}]$) and otherwise they report it as neutral. This

gives the following hypothesis: suppose we take another group of individuals and ask them to simply evaluate whether facts are neutral or lean in a particular direction. This means that the percentage of agreement between our original experimental participants and these new individuals should vary depending on information valence: both groups should agree fairly often when information is favorable however, individuals who are in our trivia game should, when restricted to 'unfavorable' facts, be more likely to say they are neutral than the group simply evaluating the facts. Again, this is not a perfect test of our hypothesis but it is a useful exercise to consider.

First, we recruited participants ($n = 86$) from AMT to judge the 'objective' nature of our facts. To do this, each participant was presented with a 'non-directional' version of each trivia statement (eg. "The temperature on day Y in X was less than Z degrees" became "The temperature on day Y in X was less or more than Z degrees.") as well as both possible facts. They then rated the facts as directional (e.g. "This fact makes me think that the temperature was more than Z") or neutral ("This fact doesn't help me evaluate this statement."). We coded these directional ratings as $+1$ if the participant said that the fact made the 'greater than' side of the statement more likely, -1 if they said that the fact made the 'less than' side more likely and 0 if they stated that the fact was neutral. We then called facts 'objectively more' if the average rating was significantly positive, 'objectively less' if the average rating was significantly negative and neutral if it is was not different from 0. Out of 20 possible facts we found that 15 were non-neutral according to Turkers. Then, considering the particular trivia statement that participants received we classified facts as 'objectively favorable' or 'objectively unfavorable.' For example, if a trivia statement was about 'X is more than Z' facts which were rated as 'objectively more' became 'objectively favorable.'

We then looked at the likelihood that a fact which was rated 'objectively favorable' was actually rated favorable by our experimental participants, this occurred 73.5% of the time. In contrast, a fact which was rated 'objectively unfavorable' was rated unfavorable by our participants only 50.4% of the time. This difference is significant at the $p < .01$ level (rank-sum test). This data suggests that one mechanism underlying our effect is the underweighting of unfavorable information.

3.9 Experiment 4: Unshrouding Effects

Existing accounts of ambiguity aversion point to another possible mechanism for our asymmetry: an aversion to ignorance. Individuals appear to care not only about their estimate of the likelihood of an event, but also about how accurate they feel their estimate is. Thus, more ambiguous situations are aversive because individuals do not feel they can accurately estimate the likelihoods of outcomes. If this is the case, then each piece of information has two effects: one on the likelihood of a winning outcome, and one on certainty. Favorable information increases both likelihood estimations and certainty (both positive effects) while unfavorable information has a negative component (decreasing likelihood) but also a positive component (increasing certainty). This mechanism, which we call "unshrouding" could also account for a part of our observed asymmetry.

In our next experiment, we test for the presence of this mechanism.

3.9.1 Methods

Two hundred participants were recruited from AMT and were presented with hypothetical pull-a-chip rounds in a within-subject design. We expanded our initial stimulus set with 6 additional randomly generated knowledge levels with X and Y given by: $(45, 30), (40, 15), (35, 60), (15, 80), (70, 15)$ and $(15, 15)$ this gave us 15 possible knowledge levels to draw from. Each participant played an $X = 50, Y = 50$ round followed by a completely ambiguous round followed by 5 additional rounds whose knowledge levels were drawn from the full set of 15.

In each round participants indicated their WTP for a ticket which won if a red chip were drawn. They also indicated their "estimate" that a red or blue chip would be drawn using an 11 point scale from -5 (Red for Sure), 0 (Either color is equally likely) to 5 (Blue for sure) in a design similar to the control in experiment 1. Finally, they were asked to indicate how "certain they felt their estimate was accurate" from a 0 (Not very certain) to 7 (Extremely certain) scale. The style of the questions were explained to participants in the instructions using the example of a coin toss (equally likely, totally certain of accuracy) and a sports game between unknown teams (equally likely, not very certain of accuracy).

See the appendix for study instructions as well as sample rounds.

3.9.2 Results

Regressing WTP for a red ticket on $numred$ and $numblue$ replicates our earlier results (table 3.11). We again find gender effects, males appear to respond much more strongly than females to favorable information (see the interaction terms).

However, we now can define two new variables of interest: $llred$ which is our likelihood score (reverse coded so higher numbers mean a higher subjective estimate of the probability of a red chip coming out) and $cert$ which is the individuals' certainty rating. First, we plot WTP as function $llred$. figure 3.3 shows a strong correlation:

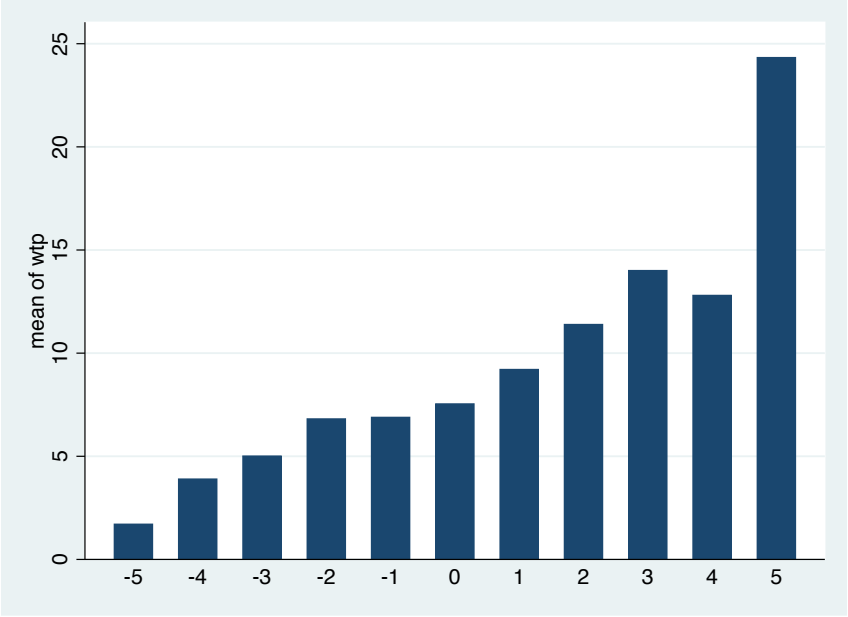

Figure 3.3: Experiment 4: WTP by reported subjective likelihood of red

Figure 3.4 shows WTP as a function of certainty for each possible level of $llred$. Overall, there is a positive effect of certainty at almost all levels of $llred$. Less than 10% of the data includes levels of $llred$ of less than -3 or greater than $+3$ – so I have dropped those. We can confirm both of these effects with a regression: regressing WTP on $llred$ and $cert$ gives positive, significant coefficients on both variables (table 3.12). Thus, both likelihood

(point estimate) and certainty (prior divergence) appear to have positive effects on WTP as the model above predicts. To make comparisons of effect sizes easier, table 3.22 in the appendix shows the same regressions performed on standardized version of the likelihood and certainty variables – both are substantial, a one deviation increase in $llred$ on average increases WTP by $\$2$ whereas a one standard deviation increase in reported certainty increases WTP by approximately $\$1$.

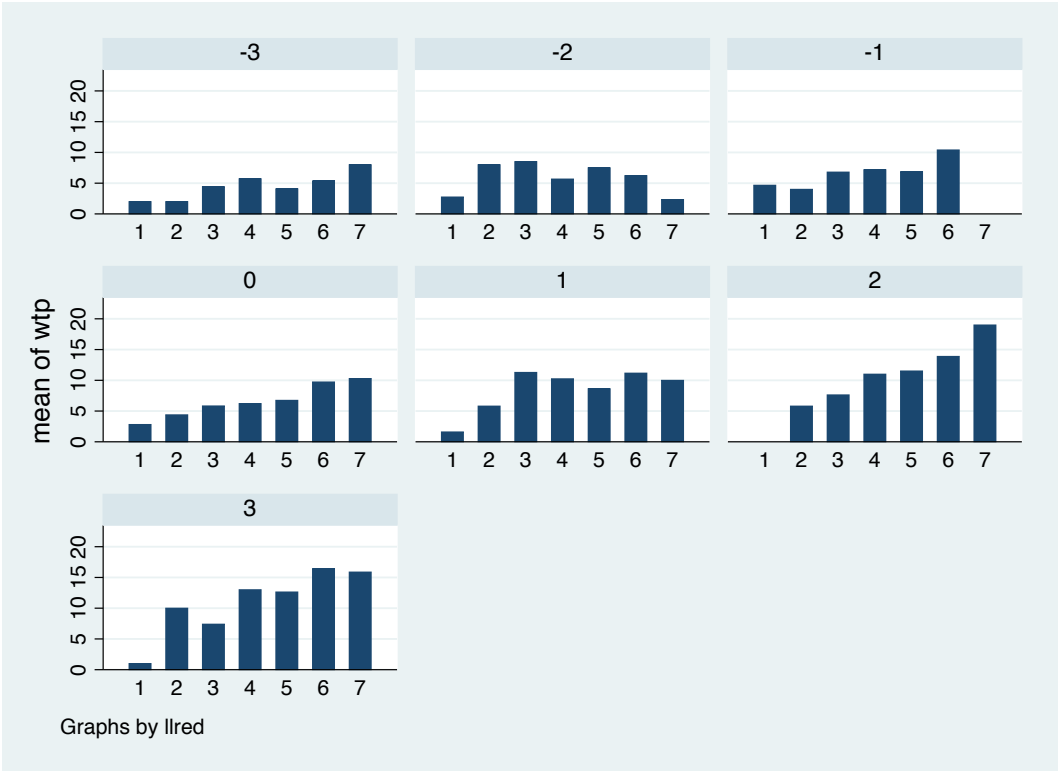

Figure 3.4: Experiment 4: WTP by reported reported certainty for various reported levels of reported likelihood of red being drawn. Each panel represents a fixed level of reported likelihood, with certainty on the X axis.

We can also see if information has the expected effects on certainty and likelihood estimates. Table 3.13 shows a regression of $llred$ (column 1) and $cert$ (column 2) on $numred$ and $numblue$. We see that information about red chips has a positive and significant effect on estimations of likelihood while the number of blue chips has a negative and significant effect. Note that there appears to be some bias towards the integration of favorable

information. Looking at the effects on likelihood estimates, the coefficients on $numred$ and $numblue$ differ in magnitudes significantly (Wald test $p < .01$) but the magnitude of the effect is on the order of 15%. However, both $numred$ and $numblue$ have significant positive effects as would be expected if our measure was capturing something like prior dispersion.

We can do a robustness check of using not simply levels of likelihood and certainty as a measure but individual *changes* in certainty. To do this, we can form new variables lld and $certd$ by differencing reported likelihood/certainty at each knowledge level with baseline likelihood/certainty in the completely unknown case. In this case, we are left with variables that are positive over 90% of the time, as it should be. We can now repeat our regression analysis by regressing WTP on lld and $certd$. Table 3.23 in the appendix shows that our effects are robust to using this analysis.

3.10 Conclusion

Our series of experiments shows that the behavioral effects of information on ambiguous valuations are very asymmetric: favorable information strongly increases valuations while unfavorable information has a much smaller effect. However, this is driven by a combination of two mechanisms: both a bias towards the integration of favorable information and also an unshrouding effect.

3.11 Tables

Table 3.8: Experiment 2 (Trivia): WTP on subjectively reported knowledge levels

	(1) wtp	(2) wtp	(3) wtp	(4) wtp
subjgood	7.293**	8.768**	9.449**	11.52**
	(1.022)	(1.371)	(1.575)	(2.197)
subjbad	-2.712**	-1.204	-1.717	-1.145
	(0.974)	(1.327)	(1.459)	(1.839)
know		0.860*	0.891*	1.145*
		(0.383)	(0.375)	(0.530)
knowXsgood		-0.692	-0.558	-0.826
		(0.501)	(0.510)	(0.634)
knowXsbad		-0.693	-0.610	-1.118
		(0.617)	(0.541)	(0.722)
ismale			-1.017	-1.725
			(1.416)	(2.072)
maleXsgood			-2.897	-2.738
			(1.925)	(2.534)
maleXsbad			0.911	1.408
			(1.971)	(2.920)
Constant	7.403**	5.520**	5.834**	4.188*
	(0.765)	(1.300)	(1.435)	(2.078)
Observations	370	370	370	370

Standard errors in parentheses clustered at participant level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.9: Experiment 3: WTP for insurance on continuous variables

	(1) insure	(2) insure	(3) (Subject FE)	(4) Tobit
numred	0.0268	-0.00522	0.0268	-0.00137
	(0.0184)	(0.0241)	(0.0195)	(0.0285)
numblue	-0.0771**	-0.0904**	-0.0771**	-0.0841**
	(0.0141)	(0.0188)	(0.0149)	(0.0174)
gender		-4.094		-3.832
		(3.221)		(2.809)
genderXred		0.0566		0.0506
		(0.0354)		(0.0415)
genderXblue		0.0235		0.0297
		(0.0276)		(0.0314)
age				-0.166+
				(0.0932)
Constant	18.21**	20.53**	18.21**	25.26**
	(1.591)	(2.546)	(0.488)	(3.795)
Observations	477	477	477	477

Standard errors in parentheses clustered at participant level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.10: Experiment 3: WTP for insurance with factor model

	(1) insure	(2) insure	(3) (Subject FE)	(4) Tobit
numred=25	-1.434**	-3.019*	-3.019*	-3.892**
	(0.509)	(1.322)	(1.399)	(1.439)
numred=50	1.340	-0.245	-0.245	-1.168
	(0.920)	(1.551)	(1.642)	(1.791)
numblue=25	-0.314	-1.906	-1.906	-2.700+
	(0.661)	(1.233)	(1.305)	(1.501)
numblue=50	-3.855**	-5.434**	-5.434**	-5.656**
	(0.707)	(1.253)	(1.326)	(1.614)
numred=25 X numblue=25		2.321	2.321	3.487
		(2.036)	(2.155)	(2.254)
numred=25 X numblue=50		2.434	2.434	2.802
		(1.658)	(1.755)	(1.967)
numred=50 X numblue=25		2.453	2.453	3.444+
		(1.555)	(1.646)	(1.776)
numred=50 X numblue=50		2.302	2.302	4.060+
		(1.719)	(1.820)	(2.072)
gender		-2.581		-1.807
		(3.035)		(2.727)
age		-0.177		-0.165+
		(0.110)		(0.0931)
Constant	18.38**	26.81**	19.43**	26.04**
	(1.620)	(4.648)	(0.928)	(3.957)
Observations	477	477	477	477

Standard errors in parentheses clustered at participant level.

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.11: Experiment 4: WTP as a function of information

	(1) wtp	(2) wtp	(3) wtp	(4) wtp
numred	0.139**	0.139**	0.0920**	0.100**
	(0.0133)	(0.0128)	(0.0146)	(0.0157)
numblue	-0.0360**	-0.0328**	-0.0323**	-0.0409**
	(0.00651)	(0.00616)	(0.00756)	(0.00925)
gender		4.080**	1.966*	1.892*
		(0.815)	(0.792)	(0.851)
age		-0.0863	-0.0864	-0.106+
		(0.0525)	(0.0525)	(0.0596)
maleXred			0.0744**	0.0748**
			(0.0232)	(0.0244)
maleXblue			-0.00118	0.00522
			(0.0115)	(0.0130)
Constant	5.088**	5.034**	6.382**	6.734**
	(0.421)	(1.747)	(1.753)	(1.948)
sigma				
Constant				7.242**
				(0.373)
Observations	1162	1162	1162	1162

Standard errors in parentheses clustered at participant level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.12: Experiment 4: WTP as a function of likelihood and certainty estimates

	(1) wtp	(2) wtp	(3) wtp
llred	1.332**	1.418**	1.068**
	(0.137)	(0.133)	(0.148)
cert	1.064**	1.047**	0.717**
	(0.157)	(0.145)	(0.158)
gender		4.315**	2.058*
		(0.799)	(0.968)
age		-0.103*	-0.101*
		(0.0505)	(0.0500)
maleXll			0.541*
			(0.238)
maleXcert			0.490+
			(0.262)
Constant	3.232**	3.728*	5.237**
	(0.620)	(1.839)	(1.625)
Observations	1162	1162	1162

Standard errors in parentheses clustered at participant level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.13: Experiment 4: Likelihood and certainty as functions of information

	(1) llred	(2) cert
numred	0.0477**	0.0296**
	(0.00213)	(0.00210)
numblue	-0.0424**	0.0243**
	(0.00176)	(0.00171)
Constant	-0.401**	3.278**
	(0.0993)	(0.143)
Observations	1162	1162

Standard errors in parentheses clustered at participant level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.14: Experiment 1 WTP using continuous variables (pull-a-chip only)

	(1) wtp	(2) wtp	(3) wtp	(4) Tobit
numred	0.154**	0.156**	0.135**	0.159**
	(0.0186)	(0.0187)	(0.0203)	(0.0225)
numblue	-0.00255	-0.00179	-0.00422	-0.0124
	(0.0121)	(0.0123)	(0.0144)	(0.0169)
gender		3.063*	1.903+	2.108+
		(1.216)	(1.107)	(1.277)
age		-0.104+	-0.104+	-0.106+
		(0.0528)	(0.0529)	(0.0577)
genderXred			0.0428	0.0338
			(0.0375)	(0.0401)
genderXblue			0.00412	0.00677
			(0.0248)	(0.0275)
Constant	5.909**	7.611**	8.219**	7.486**
	(0.616)	(1.936)	(1.846)	(2.073)
Observations	525	525	525	525

Standard errors in parentheses clustered at participant level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.15: Experiment 1 WTP using continuous variables (majority only)

	(1) wtp	(2) wtp	(3) wtp
numred	0.166**	0.170**	0.103**
	(0.0208)	(0.0208)	(0.0276)
numblue	-0.0382**	-0.0388**	-0.0115
	(0.0129)	(0.0129)	(0.0157)
gender		4.832**	3.042+
		(1.462)	(1.613)
age		-0.0267	-0.0268
		(0.0680)	(0.0685)
genderXred			0.109**
			(0.0389)
genderXblue			-0.0423+
			(0.0243)
Constant	7.475**	5.426*	6.510*
	(0.767)	(2.483)	(2.515)
Observations	550	550	550

Standard errors in parentheses clustered at participant level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.16: Experiment 1: Confidence using continuous variables

	(1) conf	(2) conf	(3) conf
numred	0.0394**	0.0362**	0.0337**
	(0.00232)	(0.00338)	(0.00480)
numblue	-0.00470*	-0.00765*	-0.00338
	(0.00229)	(0.00345)	(0.00395)
pull		-0.277	-0.275
		(0.181)	(0.183)
pullXred		0.00678	0.00741
		(0.00457)	(0.00471)
pullXblue		0.00589	0.00491
		(0.00454)	(0.00441)
gender		0.163	0.206
		(0.131)	(0.178)
age		-0.0150*	-0.0150*
		(0.00610)	(0.00609)
genderXred			0.00376
			(0.00477)
genderXblue			-0.00687
			(0.00447)
Constant	2.567**	3.087**	3.068**
	(0.0919)	(0.232)	(0.236)
Observations	899	899	899

Standard errors in parentheses clustered at participant level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.17: Experiment 1 (with choice): Confidence on continuous variables

	(1) conf	(2) conf	(3) conf
chosencolor	0.0230**	0.0260**	0.0161**
	(0.00259)	(0.00425)	(0.00580)
othercolor	-0.000956	-0.00744	0.00407
	(0.00291)	(0.00456)	(0.00604)
pull		0.445*	0.460*
		(0.213)	(0.212)
pullXchosen		-0.00612	-0.00497
		(0.00525)	(0.00523)
pullXother		0.0147**	0.0131*
		(0.00562)	(0.00566)
gender		0.0625	0.0113
		(0.187)	(0.213)
age		0.0108	0.0102
		(0.00864)	(0.00849)
genderXchosen			0.0141**
			(0.00538)
genderXother			-0.0161**
			(0.00569)
Constant	2.597**	2.048**	2.084**
	(0.108)	(0.338)	(0.343)
Observations	709	709	709

Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.18: Experiment 2 (Poker Chips - Pull-a-Chip Only): WTP on continuous variables

	(1) wtp	(2) wtp	(3) (Subject FE)	(4) Tobit
main				
numgood	0.0993**	0.0993**	0.0993**	0.109**
	(0.0137)	(0.0137)	(0.0145)	(0.0155)
numbad	-0.0123	-0.0123	-0.0123	-0.0121
	(0.00938)	(0.00940)	(0.00993)	(0.0110)
ismale		-0.453		-0.517
		(1.139)		(1.237)
age		0.112		0.121
		(0.204)		(0.214)
Constant	6.362**	4.086	6.362**	3.403
	(0.646)	(4.423)	(0.463)	(4.657)
Observations	333	333	333	333

Standard errors in parentheses clustered at participant level.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.19: Experiment 2 (Poker Chips - Majority Only): WTP on continuous variables

	(1) wtp	(2) wtp	(3) (Subject FE)	(4) Tobit
main				
numgood	0.138**	0.138**	0.138**	0.165**
	(0.0142)	(0.0142)	(0.0150)	(0.0178)
numbad	-0.0555**	-0.0555**	-0.0555**	-0.0669**
	(0.0148)	(0.0148)	(0.0157)	(0.0189)
ismale		-1.158		-1.657
		(1.177)		(1.415)
age		0.228		0.299
		(0.233)		(0.269)
Constant	6.416**	1.876	6.416**	-0.440
	(0.742)	(5.059)	(0.446)	(5.852)
Observations	333	333	333	333

Standard errors in parentheses clustered at participant level.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.20: Experiment 2 (Poker Chips): Confidence on continuous variables

	(1) conf	(2) conf	(3) conf
numgood	0.0322**	0.0364**	0.0353**
	(0.00288)	(0.00378)	(0.00421)
numbad	-0.00955**	-0.0160**	-0.0126**
	(0.00280)	(0.00389)	(0.00440)
pull		-0.192	-0.192
		(0.167)	(0.167)
pullXgood		-0.00837*	-0.00836*
		(0.00402)	(0.00403)
pullXbad		0.0130**	0.0130**
		(0.00302)	(0.00302)
ismale		0.176	0.335
		(0.200)	(0.285)
age		-0.0253	-0.0252
		(0.0350)	(0.0350)
maleXgood			0.00279
			(0.00605)
maleXbad			-0.00915
			(0.00548)
Constant	2.481**	3.062**	3.002**
	(0.135)	(0.776)	(0.772)
Observations	665	665	665

Standard errors in parentheses clustered at participant level.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.21: Experiment 2 (Trivia): WTP on subjectively reported knowledge levels

	(1) conf	(2) conf	(3) conf
subjgood	1.422**	1.664**	1.836**
	(0.258)	(0.407)	(0.421)
subjbad	0.246	0.163	0.0786
	(0.183)	(0.299)	(0.316)
know		0.661**	0.656**
		(0.0853)	(0.0818)
knowXsgood		-0.174	-0.137
		(0.111)	(0.100)
knowXsbad		-0.0228	0.00153
		(0.105)	(0.111)
ismale			-0.0186
			(0.238)
maleXsgood			-0.730+
			(0.390)
maleXsbad			0.0965
			(0.288)
Constant	2.653**	1.195**	1.210**
	(0.193)	(0.243)	(0.276)
Observations	370	370	370

Standard errors in parentheses clustered at participant level

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.22: Experiment 4: WTP as a function of likelihood and certainty estimates (standardized)

	(1) wtp	(2) wtp	(3) wtp
llred(std)	2.680**	2.854**	2.150**
	(0.277)	(0.269)	(0.299)
cert(std)	1.856**	1.826**	1.250**
	(0.275)	(0.252)	(0.275)
gender		4.315**	4.293**
		(0.799)	(0.799)
age		-0.103*	-0.101*
		(0.0505)	(0.0500)
maleXlls			1.088*
			(0.480)
maleXcerts			0.855+
			(0.457)
Constant	8.043**	8.436**	8.435**
	(0.459)	(1.675)	(1.667)
Observations	1162	1162	1162

Standard errors in parentheses clustered at participant level

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 3.23: Experiment 4: WTP as a function of likelihood and certainty estimates (differences)

	(1) wtp	(2) wtp	(3) Tobit
lld	1.358**	1.263**	1.360**
	(0.159)	(0.153)	(0.166)
certd	0.467*	0.451**	0.447*
	(0.181)	(0.172)	(0.181)
gender		3.562**	3.654**
		(0.816)	(0.870)
age		-0.0620	-0.0795
		(0.0552)	(0.0623)
Constant	7.093**	6.777**	7.075**
	(0.508)	(1.809)	(2.009)
Observations	1162	1162	1162

Standard errors in parentheses clustered at participant level
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Chapter 4

How to Commit (If You Must)

4.1 Brief Summary

This paper studies how dual-self (Fudenberg and Levine (2006)) decision-makers can use commitment technologies to combat temptation and implement long-run optimal actions. I consider two types of such technologies: carrot contracts (rewards for 'good' behavior financed by borrowing from future consumption) and stick contracts (self imposed fines for 'bad' behavior). Both types of contracts can simulate binding commitment when it is not available and thus offers of such contracts make DMs better off. I show that the exact properties of optimal carrots and sticks depend crucially on the short-run discount rate and the time it takes for the contract to 'kick in.' Finally, I compare the welfare implications of these contracts and show that dual-self decision-makers strictly prefer to use carrots instead of either sticks or binding commitments. This is for several reasons: sticks are highly vulnerable to trembles (while carrots are not), sticks and binding commitments create a temptation to cancel them (while carrots do not), and finally carrots allow easy tradeoffs between commitment and flexibility (while sticks and binding commitments do not).

4.2 Introduction

Many of us would like to exercise, work efficiently and stay away from our bad habits yet we often find ourselves skipping a daily run, looking at funny cat videos on the internet, smoking another cigarette or eating another cupcake. Such mismatches between what we would like to do tomorrow and what we actually end up doing create a demand for technologies to help individuals implement their normative goals. Existing literature[1] shows that this demand for commitment does exist but there is little theoretical work on what forms optimal commitment technologies would take. This paper begins to bridge this gap.

The first step to figuring out optimal commitment technologies is learning what mechanism generates the problem in the first place. Work in psychology and neuroscience tends to focus on decisions as an interplay between an automatic and a controlled process (Kahneman (2003)). Recent work in social psychology makes this model more sharp by positing that the controlled cognitive process uses a limited, costly resource to operate. A standard template in such experiments involves subjects performing a 'resource depleting task' (controlling attention, suppressing emotion, solving math problems) or a control task followed by a second resource depleting task. Subjects who are had to perform the depleting task do worse on the second task than controls (see Muraven and Baumeister (2000) for a survey). Another set of experiments involve individuals making choices under cognitive load (for example, subjects are asked to remember a 7 digit number), these individuals then act more impulsively than controls (eg. by choosing more unhealthy foods to eat as in Shiv and Fedorikhin (1999)). In such a model our difficulties at the desert tray come from an interplay of automatic impulses compelling us to eat, and a conscious use of mental resources not to give in.[2]

This paper uses a particular economic model of this process: the dual-self model of Fudenberg and Levine (2006) (henceforth FL). This model is a specific case of a larger

[1] Ashraf et al. (2006), Kaur et al. (2010), Ariely and Wertenbroch (2002) consider commitment in the field, Houser et al. (2010) considers commitment in the lab. Bryan et al. (2010) survey the existing research.

[2] Work such as Hare et al. (2009) in neuroeconomics points to a possible neural algorithm for this model in which control networks in the dorsolateral prefrontal cortex modulate 'overreactions' by the brain's reward system.

set of costly self-control models.[3] In such models decisions are a compromise between a 'temptation' ranking and a 'normative' ranking, with the DM balancing a desire to choose according to his normative preference with a self-control cost of deviating from the temptation ranking.

The FL model imposes specific restrictions on where the disagreement between the temptation and normative preferences comes from: the DM uses a standard time-separable utility function to evaluate consumption streams (or dynamic plans) but the temptation or automatic process discounts the future at a much sharper rate than the DM would like. Thus the DM is tempted to behave impulsively and must use self-control to choose long-run rewards. Because of this structure FL refer to the automatic impulse as the short-run (SR) self and the cognitive control process as the long-run (LR) self.[4] Note that this is a different type of model than those studied in the literature on hyperbolic/quasi-hyperbolic discounting (eg. Laibson (1997), Ainslie and Haslam (1992)). In those models the DM's problems come from the fact that rankings of alternatives change in different periods. In the FL model both the SR and LR self are perfectly time consistent so the tension comes from multiple preferences within a period rather than multiple preferences between periods. The purpose of this paper is not to provide a clean test to differentiate these models, rather it is to look at commitment behavior with the FL model. However, all results that could never hold under a time-inconsistent framework are flagged as such.

The FL model generates a huge demand for commitment and this paper considers two types of technologies that could be available to a DM facing temptation. The first are stick contracts that levy a fine on the DM when he gives in to temptation.[5] The second are carrots

[3]Gul and Pesendorfer (2001), Dekel et al. (2009), Gul and Pesendorfer (2004), Noor and Takeoka (2010) are highly visible examples in which self-control costs are variable and depend on the amount of adjustment the control process needs to do. Benhabib and Bisin (2005) consider the case where the control process is treated as having a fixed cost to activate.

[4]The analysis of Fudenberg and Levine (2006) considers the case where the SR self is perfectly myopic, Fudenberg and Levine (2012) extends to the case where the SR self can have some degree of patience.

[5]A market version of such contracts is provided by the website StickK.com that allows individuals to set measurable goals, punishments and a referee. If an individual does not accomplish his goal, as reported by the referee, the website will automatically charge the individual's credit card a donation to an 'anti-charity' of his choice (for example, a life-long Democrat may choose to donate to finance the George W. Bush Presidential Library.)

that reward a DM who takes normatively good actions. In this paper, carrots are financed by the intertemporal substitution of future consumption to the present conditional on the DM resisting a temptation. The main results show that both of these types of contracts can only be welfare improving for a DM if they change the nature of the temptation he faces - i.e. the SR self's optimal action. The logic is one of revealed preference combined with the economic idea that self-control is treated as a cost. If a DM gives in to a temptation, this is because the self-control required to resist it was too expensive. If a commitment technology does not physically remove the tempting option, then it must remove the temptation associated with that option because if it does not, its ultimate effect is only to make the DM exert the self-control that he didn't find optimal to exert in the first place. This means that if the source of the temptation is sharp discounting, both types of contracts must have one of two features: either their effects must be close in time to the choice or they must be particularly large. In fact, as the paper shows later, their size increases exponentially with the delay between action and punishment.

The natural question to ask given these results is whether carrots, sticks or binding commitments are favored by the dual-self DMs. The final set of results shows that carrots are preferred for several reasons. First, sticks will implement a punishment if the DM trembles and executes unintended actions with small probability whereas a tremble in the presence of a carrot makes the DM no worse off than before the contract. Second, sticks and binding commitments require self-control to implement in the first place and if the DM receives opportunities to cancel the contract, the cancellation acts as an additional temptation and source of self-control problems in the case of sticks and binding commitments but not carrots. Finally, the DM may have a desire for flexibility. If the temptation's size is stochastic optimal carrots allow the DM to retain the flexibility choose the temptation when it is LR optimal while sticks and binding commitments do not.

The results in this paper are related to those of Ali (2011) who shows that carrots are advantageous for DMs seeking flexibility in a planner-doer model of self-control in which the planner attempts to learn about the doer's preferences. From a theoretical perspective, this paper together with the results of Ali (2011) indicate that there is much to be learned by adding dynamic behavior into models of self-control. More practically they also indicate that devices beyond binding commitments and self-punishing technologies may be useful

ways of dealing with self-control problems.

4.3 The Basic Model

We now introduce the basic dual-self model used in Fudenberg and Levine (2012) in continuous time. All formal proofs of results are relegated to the appendix.

The DM begins at time $t = 1$ and faces a 'simple temptation.' He chooses an action from the set $\{T(ake), R(exist)\}$. If he chooses T he gains a benefit b at that moment, however, at $t = 2$ he takes a loss of 1. If he chooses R, he gains no payoffs but suffers no loses later. Taking a temptation means that the DM takes a payoff now for a loss later while R keeps payoffs constant between periods.

The DM is split into two 'selves' that interact to make decisions. The long-run (LR) self discounts the future with an instantaneous discount rate of ρ (which we take to be 0 for simplicity) and the short-run (SR) self discounts at a faster rate λ. The intuition behind how the selves interact is as follows: when the DM faces a choice, the SR self 'suggests' to take the action that maximizes SR utility, the LR self can then choose to either go with this suggestion or to change to a different action. However, if the LR self wishes to change course he must pay a self-control cost to do so. This cost is proportional to the amount of utility the SR self must give up. Thus, this model represents in a simple, stylized way an interaction between 'automatic' or 'affective' processes that drive individuals toward immediate rewards (the SR self) and cognitive processes (the LR self) that can be invoked to control them and that become more costly as temptations become larger. For the rest of this exposition, as has been common in the literature, the LR self's preferences will be used whenever welfare metrics for the DM are discussed.

We now formalize the discussion above: suppose that the DM faces a set of consumption streams A, the LR self's utility from an action x is given by

$$u_{LR}(x) - SC(x)$$

where $u_{LR}(x)$ is the discounted value of x using the discount rate ρ and $SC(x)$ is the self-control cost the LR self must pay to implement action x. The self-control cost of taking the

SR self's preferred action x^*_{SR} is 0 and the cost of self control from choosing a different action is given by

$$SC(x) = \psi(u_{SR}(x^*_{SR}) - u_{SR}(x))$$

where $u_{SR}(x)$ is the discounted value of x using the SR discount rate λ.

Thus, the LR self will use self control whenever the benefit from choosing x instead of x^*_{SR} exceeds the self-control cost required to do so. In the analysis that follows, ψ will be assumed to be a linear function as in Fudenberg and Levine (2006) but the main results in this paper are robust to using a different functional form.

Note that the LR self is perfectly time consistent but does display demand for binding commitment as the mere existence of tempting options creates a potentially costly conflict. Note as well that unlike in models of time-inconsistency, the DM may want to remove options from future choice sets even when he knows for sure that he will not take them.

Applying this model to the simple temptation problem, this interaction is played out in the following way: if R is chosen, the SR self gains a utility of $u_{SR}(R) = 0$ while if T is chosen the SR self gains utility

$$u_{SR}(T) = b - e^{-\lambda}.$$

Suppose that $b - e^{-\lambda} > 0$ so the SR self prefers to take the temptation. This means the LR self gets utility 0 from R and gets utility from choosing T given by

$$b - 1 - SC(T).$$

To motivate the problem, we take $b < 1$ and $-\psi(b - e^{-\lambda}) < b - 1$. Without any intervention we have that the SR self prefers to take, the LR self prefers to resist but will not do so given the size of self-control costs.

4.3.1 Sticks

We now consider how the DM can implement the long-run optimal action in the absence of perfectly binding commitment devices. First we consider how a DM is able to use contracts that assign a penalty of size k to a choice of T (sticks). This penalty is delivered at time $t' \in [1, 2]$. Intuitively, one can think of this as how long it takes for the contract

enforcement mechanism to observe the agent's action and implement the punishment. Several of the results consider what happens when $t' > 1$ but setting $t' = 1$ does not change other results. Figure 4.1 shows the full timeline.

We can now turn to existing sticks for intuition: alcoholics sometimes take the drug Antabuse. This drug stops the body from correctly metabolizing alcohol - the result makes it so that taking even a small drink of alcohol results almost immediately into an experience similar to a severe hangover. Control-It! is a foul tasting (but completely safe) liquid that individuals who are attempting to quit biting their nails apply to them. Individuals trying to lose weight take the drug Xenecal (available over-the-counter as Alli in the United States). This drug acts in a similar manner to Antabuse by making the eating of fats a highly unpleasant experience. In each of these examples punishments come almost immediately after the bad behavior in question and so t' is very close to 1. However, if we consider StickK.com, punishments are only realized after the DM's referee reports that he has chosen T. In this case, it seems reasonable to think that t' will be bounded away from $t = 1$ by a non-trivial amount.

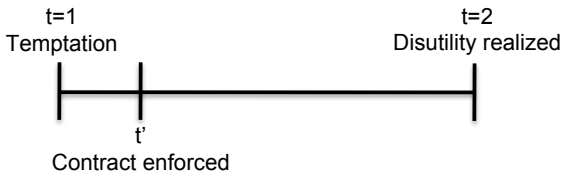

Figure 4.1: Timeline of the base problem.

In this setup LR utility at the time of choice of choosing to submit to temptation given a stick of size k is given by

$$u_{LR}(T, k) = b - k - 1$$

and the SR utility of the same is given by

$$u_{SR}(T, k) = b - e^{-\lambda t'} k - e^{-\lambda}.$$

To make notation simpler, for an action $\sigma \in \{T, R\}$ and contract of size k we define $\eta(\sigma, k)$ to be the self-control cost associated with taking that action as introduced above. Let $\sigma^*(k)$

be the action that maximizes the LR self's utility net of self-control costs (and thus the DM's welfare). We denote the welfare by

$$W(\sigma, k) = u_{LR}(\sigma, k) - \eta(\sigma, k).$$

We can now discuss the effects of the contracts on the DM's behavior and welfare. First, define:

Definition 2. *Say that a stick **implements resisting** if $\sigma^*(k) = R$. Let $\mathcal{K}^R$ be the set of sticks that implement resisting.*

We now turn to defining optimal sticks:

Definition 3. *The set of **optimal sticks** is defined by*

$$\mathcal{K}^* := \{k \in \mathbb{R}_+ \mid W(\sigma^*(k), k) \geq W(\sigma^*(k'), k')$$

for any $k' \in \mathbb{R}_+$.

The next proposition concerns the perfectly myopic SR selves studied in Fudenberg and Levine (2006).

Proposition 4. *Suppose the SR self completely discounts all future payoffs and $t' > 1$, then $\mathcal{K}^R \neq \emptyset$ but $\mathcal{K}^* = \{0\}$.*

This result states that when the SR self is perfectly myopic, the DM is always made worse off with a stick that does not punish him immediately when he succumbs to temptation. This means that under such circumstances, the DM will never take a stick when offered. The intuition for the proposition comes from the rational way in which self-control is treated in the FL model. We know that under the condition $k = 0$ the cost of self-control to implement resisting was larger than the foregone gains to the LR self. Since at time $t = 0$ the whole of the contract is in the future a perfectly myopic SR self is completely unaffected by the imposition of the contract; the only channel that remains for the contract to operate through is making the future consequences of the temptation so unattractive that the LR self will choose to exercise self-control and avoid the extra punishment. Thus the existence of the contract represents a welfare decrease for the LR self. Allowing for SR discount rates below 100% recovers the existence of optimal sticks:

Proposition 5. *Suppose that the SR self discounts at a positive but not infinite rate, then* $\mathcal{K}^* \subsetneq \mathcal{K}^R$. *Moreover*

$$\mathcal{K}^* = \{k \geq k^* = \frac{b - e^{-\lambda}}{e^{-\lambda t'}}\}$$

and so is independent of ψ *and LR discounting.*

This proof of the proposition shows that for an FL DM works in a very particular way. Simply implementing resisting is not enough to make an optimal contract, rather the possible loss from the contract must be large enough to change the optimal course of action for the SR self. Note that it is possible to refine this set further using the intuition that there is no incentive for the DM to take a contract beyond the size of k^* as a larger stick does not change any behavior on the equilibrium path but increases possible losses to the DM in the case of a tremble. This intuition is developed in the Appendix.

There are two other points to be taken from this analysis. First:

Corollary 1. *For any value of* λ *self-control is never exercised with any optimal stick.*

This implies that optimal sticks can simulate perfectly binding commitments and thus are a useful device for dual-self DMs. Additionally, in the dual-self model, sticks have one additional property:

Corollary 2. *For any value of* λ *there exists an open set* $(\underline{k}, \overline{k})$ *such that* $k \in (\underline{k}, \overline{k})$ *implements resisting but does not improve welfare over* $k = 0$.

Note that this is would be impossible in a model of time-inconsistency (assuming that we used the common welfare criterion of the $t = 0$ self) as in such a model any commitment device that implemented R would necessarily be welfare increasing. In addition, this interval can, in general, be quite large as $\underline{k}$ is a constant and $\overline{k}$ strictly increases in λ.

The next proposition characterizes a property that all optimal sticks must display:

Corollary 3. *The lower bound for optimal sticks* k^* *grows exponentially as the enforcement time,* t' *moves away from* 1. *The speed of this growth is proportional to the SR discount rate* λ.

While in a perfect world this proposition poses no problems, in a world where DMs can tremble to $Take$ with a small probability there is a completely different story. When

contracts can be taken that employ punishment almost immediately, their effects do not have to be extraordinarily large but as we push the time of resolution backward and the contract grows exponentially we run into problems.

To appreciate the impact of this growth, consider a case where $b = .5$ and that the SR self discounts at a rate of 50% per day with $t = 2$ approximately 1 month later.[6] If t' is three days later the optimal stick size is approximately $k = 4$ – this is 8 times the gain from the contract. If t' is a week later then the optimal contract has a punishment size of 64 times the utility gain from implementing the normatively superior action. In this case, even a 2% chance of the DM trembling to adhere to the long-run optimal action makes the contract not worth it. A similar result can be recovered if we require simply that the stick is welfare improving and not necessarily optimal.

4.3.2 Carrots

We now consider a different type of commitment contract, a carrot. In this type of contract the DM receives a reward of utility size r if he takes the long-run optimal action. For simplicity, we assume that this reward is financed by borrowing from future consumption. For example, a DM may commit to 'buying themselves something nice' if they manage to finish a particular project on time. Additionally, to make the comparison between the costless sticks and carrots possible we assume that the LR preference is indifferent between consuming r today and the discounted future value of r.[7]

The timeline with carrots is identical to that of the one with sticks: the carrot is administered at time $t' \in [1, 2]$ if the long-run optimal action R is taken. If the DM chooses T then the carrot is not activated and the future consumption stream of the DM is untouched.

[6] Pinning down SR discount rates is an important but difficult empirical challenge. For example, McClure et al. (2007) consider the case of primary rewards (juice for thirsty individuals) and show that the discount horizon for the 'short-run' is about $\sim 60\%$ *per* 25 *minutes* in their experiment. However in many other experiments (including any ones that involve monetary payouts) individuals appear to react impulsively to short run rewards that they will not receive until at least the end of the experiment. Clearly, SR discounting must display some sort of context dependence.

[7] If this neutrality isn't satisfied then the intertemporal loss from changing consumption can be viewed as a shadow cost of using the carrot contract even in the absence of other costs. Thus, the DM will only take a carrot if the intertemporal loss is less than $1 - b$. The appendix discusses in more detail the magnitude of these shadow costs.

We maintain identical notation from the last section and define the set of resistance implementing carrots $\mathcal{R}^R$ and the set of optimal carrots $\mathcal{R}^*$ analogously.

In the case of perfectly myopic SR selves we get an analogue to proposition 4 for carrots:

Proposition 6. *Suppose the SR self completely discounts all future payoffs and $t' > 1$, then $\mathcal{R}^R = \emptyset$ and $0 \in \mathcal{R}^*$*

Thus carrots are never useful when the SR self perfectly discounts the future. However, when the SR self has non-negligible valuation of future consumption carrots become useful:

Proposition 7. *Suppose that the SR self discounts at a positive but not infinite rate, then $\mathcal{R}^* \subsetneq \mathcal{R}^R$. Moreover*

$$\mathcal{R}^* = \{r \geq r^* = \frac{b - e^{-\lambda}}{e^{-\lambda t'}}\}$$

and so is independent of ψ and LR discounting.

Finally, just like sticks, carrots are able to perfectly simulate binding commitment:

Corollary 4. *For any value of λ self-control is never exercised with any optimal carrot.*

And finally just like sticks:

Corollary 5. *The lower bound for optimal carrots r^* grows exponentially as the enforcement time, t' moves away from 1. The speed of this growth is proportional to the SR discount rate λ.*

Thus, sticks and carrots are useful replacements for DMs seeking binding, but unavailable, commitment devices.

4.4 Comparison

We now turn to comparing the welfare effects of carrots to sticks. Notice that if we assume that the DM trembles to the unintended action with probability ϵ, carrots mechanically gain an advantage over sticks because they provide no downside risk. We now consider two more factors that could be at play in real commitment decisions. First, we add

the presence of stochastic opportunities to 'call off' a commitment. Second, we consider the role of flexibility in commitment. We show that each of these factors weigh the welfare scale towards carrots and away from both sticks and binding commitments.

We consider a DM who faces the simple temptation above at time $t = 1$ but, at $t = 0$, has a chance to choose a commitment device from the menu $\{r^*, k^*, s, N\}$ where N is no commitment, k^* is the optimal stick, r^* is the optimal carrot and s is the binding contract that removes T from the DM's menu. We assume that the SR self is not perfectly myopic (thus k^* and r^* can simulate binding commitment even if $t' > 0$).[8] This is the key assumption that drives the revision result but not the flexibility result.

Furthermore we assume between $t = 0$ and $t = 1$ the DM lives in continuous time and receives opportunities to revise his choice and costlessly replace any current commitment with N. These opportunities arrive as a Poisson process with arrival rate $\mu \in [0, \infty)$. Figure 4.2 shows the structure of the problem.

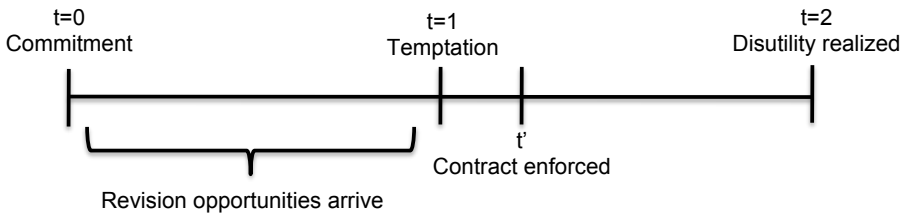

Figure 4.2: Timeline of the expanded problem.

To solve the problem the DM faces, we must specify how the SR self evaluates future actions at $t = 1$ from a $t = 0$ perspective. Here we follow the assumption in Fudenberg and Levine (2012) that the SR self is strategically naive and assumes that all future decisions will include no exercise of self-control by the LR self.[9] We get the following characterization of the DM's behavior:

[8]Fudenberg and Levine (2012) give many reasons for assuming that the SR preferences display some degree of patience. Rather than recapitulate their arguments, we point the interested reader to that paper.

[9]Note that we have already shown that no self-control is ever exercised at optimal carrots/sticks so the main results in this section could still be derived under the assumption that SR self is strategically sophisticated.

Proposition 8. *For any value of μ, the DM strictly prefers choosing the optimal carrot at time 0 to any other option. Furthermore, there exists $\overline{\mu}$ such that if $\mu > \overline{\mu}$ the DM strictly prefers no commitment to taking the optimal stick or the binding commitment device.*

The intuition for the proposition comes from how sticks and carrots simulate binding commitment in different ways. Optimal sticks and carrots both make the SR self indifferent between T and R, however they do so in different ways. The stick *lowers* the utility of the temptation for the SR self while the carrot *raises* the utility of resisting. This means that the SR self is made worse off in the future by either binding commitment or the use of a stick but is made weakly better off by the use of a carrot. The SR self isn't perfectly myopic and so now prefers carrots and no commitment to sticks and binding commitments. Because of these SR preferences taking a stick or binding commitment requires self control by the LR self when N is in the choice set, but taking a carrot does not.

The same logic holds during revision opportunities: canceling a stick or binding commitment is itself a temptation that the DM must use self-control to resist. Thus, if the DM expects many revision opportunities before $t = 1$ he may find that the expected self-control of taking a stick or binding commitment cost exceeds the benefit. Contrarily, canceling a carrot is not tempting and so the presence of revision opportunities does not affect the DM's valuation.

4.4.1 Flexibility

Another reason for the preference towards carrots over sticks comes from a desire for flexibility. We can expand the model above to one where b (the temptation) payoff is not a single number but is in fact stochastic and ex-ante unknown. Suppose that b is distributed on $[0, b_{max}]$. We suppose this distribution is well behaved and we call the PDF $f(\cdot)$ and the CDF $F(\cdot)$.[10] We now ask how demand for sticks, carrots and binding commitments changes in this new situation. As above, suppose that sticks/carrots/binding commitments are taken at time 0 and the DM is aware of the distribution of b at time 1. We set the revision parameter to 0 to isolate flexibility motives.

[10] We do not require full support assumptions.

We now compute the DM's expected welfare from being uncommitted, then compare that to the DM's welfare from having a binding commitment, a stick contract or a carrot contract. Note that when the DM is uncommitted due to continuity of the LR welfare in b we can break up the interval $[0, b_{max}]$ into several components. We know there exists a subinterval $[0, \underline{b}]$ in which the DM resists the temptation and exercises no self-control (as the SR self does not prefer T to R). We also know that there exists another subinterval $[\underline{b}, \overline{b}]$ such that for any value of b in this interval, the DM resists the temptation (but pays a self-control cost to do so). In the interval $[\overline{b}, 1]$ the DM does not resist the temptation (but is made worse off), and finally, in the interval $[1, b_{max}]$ the DM chooses T and because it is LR superior to R.

We can compute the borders for these intervals simply: the first bound is the SR self's indifference point $\underline{b} = e^{-\lambda}$ and the second is the level of b at which the DM is indifferent between taking and resisting

$$\overline{b} = \frac{1 + \psi e^{-\lambda}}{1 + \psi}.$$

We can compute that the DM is made better off under a binding commitment that under no commitment if and only if the following condition holds:

$$\int_{\underline{b}}^{\overline{b}} \psi(b - e^{-\lambda}) dF + \int_{\overline{b}}^{1} (b - 1) dF \geq \int_{1}^{b_{max}} (b - 1) dF.$$

The intuition behind this condition is the following: the left hand side of the inequality represents the gains from commitment. If b falls in the interval $\underline{b}$ to $\overline{b}$ and he were uncommitted, the DM would have to exercise self-control. If b falls in the interval $\overline{b}$ to 1 the DM would give in to the temptation. Having a binding commitment removes these two problems. However, it comes at a cost. If b falls in the interval 1 to b_{max} the DM would like, from an LR perspective, to chose T. Thus, in this case a binding commitment causes the DM to lose potential payoff.

Now, suppose we also allow the DM to choose from a full menu of sticks and carrots (that is, any values of r or k). We get the following proposition:

Proposition 9. *For any distribution of temptations $F(\cdot)$ the DM always weakly prefers sticks to binding commitments. Moreover, there exist distributions such that this preference*

is strict. Additionally, for $F(\cdot)$ that puts positive weight on any $b > 1$ there exists a carrot such that DM is always strictly better off under that carrot than under any stick or binding commitment.

This proposition comes from the intuitions developed in the last section. First, sticks are at least as valuable as binding commitments because a very large k can always exactly simulate the binding commitment. This can also hold strictly: the proof shows an example in which sticks yield a strictly higher welfare than binding commitments.

The final part of the proposition comes from the fact carrots have no downside: if the DM chooses T under a carrot contract he simply foregoes his separate reward for R, however once a DM has chosen a stick he must lose payoff whenever he chooses T. Thus, carrots allow for maximum flexibility and are strictly preferred by the DM whenever flexibility is possible.

4.5 Conclusion

Models of self-control problems all predict a large demand for commitment in our daily lives and indeed these technologies are all around us. Economists have studied binding financial commitment devices such as Christmas Clubs (Loewenstein and Thaler (1989)) or the use of illiquid assets (Laibson (1997)). In the digital world, the Apple Mac App store contains apps such as Concentrate or Self-Control that allow the user to turn off access to certain websides (eg. Facebook) during work hours. However, though these devices exist, the predicted demand of commitment seems to far outstrip the much more limited supply.

Existing literature gives one reason for this dearth: a lack of sophistication (DellaVigna and Malmendier (2006)) on the part of DMs. On the other hand, others have argued that combined with learning, an initial a lack of sophistication can only lead to *over*commitment in the long run (Ali (2011)). The results in this paper show (as others, e.g. Bryan et al. (2010), have informally stated) that there may be many other good reasons for individuals not to want binding commitments or sticks even when they know that they will be faced with temptations. In addition, at least anecdotally, it seems that many individuals do use a form of carrots to motivate themselves to fight temptations such as procrastination (for

example, agreeing to go out to a nice meal with friends conditional on finishing a work project) so actually these types of commitments may be quite prevalent and understudied.

Many interesting issues remain in the study of self-control – especially in the domain of dealing with self-control problems. This paper is meant to be a first step in applying the existing economic theories to both the understanding how individuals deal with self-control problems and the task of designing useful mechanisms that individuals could use to stick to their long-run optimal plans.

4.6 Appendix 1: Strictly Optimal Sticks

Suppose that we consider a world where the DM is restricted to trembling to the action T with probability ϵ. Assume that this tremble incurs no self-control cost (the next section of the appendix discusses in more technical detail why we use this formulation instead of a formulation where the control cost is allowed to depend on the tremble probability) so that a DM who trembles to T receives final LR utility $b - k - 1$. Let

$$W_\epsilon(\sigma^*(k), k) = (1 - \epsilon)W(\sigma^*(k), k) + \epsilon(b - k - 1).$$

This lets us define a stronger notion of optimality for stick contracts.

Definition 4. *Say that k is **a strongly optimal KM contract** if for any $k' \in \mathbb{R}_+ \setminus k$ there exists a sequence $\epsilon_n \to 0$ such that*

$$W_{\epsilon_n}(\sigma^*(k), k) > W_{\epsilon_n}(\sigma^*(k'), k) \text{ for each } n \in \mathbb{N}.$$

Let $\mathcal{K}^{**}$ be the set of strongly optimal KM contracts.

The logic behind this definition is a sort of purification argument similar to that employed in many game-theoretic refinements. The next proposition shows that the definition has bite as a refinement of the optimal stick set:

Proposition 10. *For an FL DM with $\lambda < \infty$ we have that $\mathcal{K}^{**}$ is a single point given by*

$$k^* = \frac{b - \gamma}{e^{-\lambda t'}}.$$

4.7 Appendix 2: Mixed Strategies

We may want to model not exogenously specified trembles but consider DMs who make choices that are not simply single elements of some choice set but probability distributions over the choice set. To analyze such choices, we need to make assumptions on how the choice of probabilistic actions influences the self-control costs of the LR self. In this section we will relax the linearity assumption made in the body of the paper and let ψ be a general weakly convex, smooth function with $\psi(0) = 0$.

For notation, let $\mathcal{M}(A)$ be the set of probability distributions over a finite set of alternatives $A \subset \mathcal{X}$ with generic element σ. Let $\sigma(y)$ be the probability that σ assigns to $y \in A$. There are two main ways to consider:

Definition 5. *Fix $\sigma \in \mathcal{M}(A)$, the **ex-ante expected self-control cost** denoted $\zeta(\sigma, A)$ of σ is given by*

$$\zeta(\sigma, A) = \psi(\overline{u}_{SR}(A) - \sum_{y \in A} \sigma(y) U_{SR}(y)).$$

The intuition behind this formulation is that the LR self actually chooses a probability distribution over elements of A, pays the self control cost for the distribution and then realizes some draw from the distribution. The other formulation is as follows:

Definition 6. *Fix $\sigma \in \mathcal{M}(A)$, the **ex-post expected self-control cost** denoted $\eta(\sigma, A)$ of σ is given by*

$$\eta(\sigma, A) = \sum_{y \in A} \sigma(y) \psi(\overline{u}_{SR}(A) - U_{SR}(y))$$

In this formulation the randomization is viewed as one over the LR self's control actions. Here the LR self uses the randomizing device to select an action but pays self-control costs to execute it as in the standard definition.

For simplicity we now assume that we have for all $A \subset \mathcal{X}$ and $y \neq y' \in A$

$$U_{LR}(y) - \psi(\overline{u}_{SR}(A) - U_{SR}(y)) \neq U_{LR}(y') - \psi(\overline{u}_{SR}(A) - U_{SR}(y'))$$

so the LR self is never indifferent between any two options when self-control costs are taken into account. Then fix $A \subset \mathcal{X}$ and let σ^*_η be the solution to

$$\max_{\sigma \in \mathcal{M}(A)} \sum_{x \in A} \sigma(x) U_{LR}(x) - \eta(\sigma, A)$$

and σ^*_ζ be the solution to the same problem using the ex-ante self-control formulation.

The choice of formulation is not without consequences:

Proposition 11. *For any $A \subset \mathcal{X}$ we have that $\sigma^*_\eta(x) \in \{0, 1\}$ for any $x \in A$.*

Proof. Recall that the LR utility of choosing mixed strategy σ is given by

$$\sum_{y \in A} \sigma(y) U_{LR} - \sum_{y \in A} \sigma(y) \psi(U_{SR}(x^*(A)) - U_{SR}(y)).$$

But this is just

$$\sum_{y \in A} (\sigma(y) U_{LR} - \sigma(y) \psi(U_{SR}(x^*(A)) - U_{SR}(y))).$$

By assumption of no indifference there exists a unique $y \in A$ to maximize this. □

Thus under the ex-post formulation a DM who is not constrained to choose a distribution with full support always makes a deterministic choice. This is not the case for the ex-ante formulation. Consider an example where utility is linear and the choice set A is given by the consumption streams $x = (1, 0)$ and $y = (0, 2)$. Suppose further that $\gamma = 0$, $\delta = 1$ and $\psi(z) = az^2$. The LR utility of choosing a distribution $\sigma \in \mathcal{M}(A)$ is given by

$$\sum_{x \in A} \sigma(x) U_{LR}(x) - \zeta(\sigma, A)$$

that in this case is

$$2(\sigma(y)) + (1 - \sigma(y)) - a(1 - (1 - \sigma(y)))^2.$$

It can be readily checked that the optimal solution sets

$$\sigma(y) = \frac{3}{2a}$$

for values of $a > \frac{3}{2}$ and thus is interior. However, this effect is a result of the choice of self-control cost function:

Proposition 12. *If ψ is a linear function then*

$$\eta(x, A) = \zeta(x, A)$$

for any choice of $A \subset \mathcal{X}$ and $x \in A$.

If we are not restricting ψ to be linear and since the choice of non-degenerate randomizations in basic maximization problems seems to be intuitively unappealing so the ex-post self-control formulation seems more reasonable.

4.8 Appendix 3: Proofs of Propositions

We first prove an auxiliary lemma that makes the following results much simpler.

Lemma 1. *The function $W(\sigma^*(k), k)$ is continuous in k.*

Proof of Lemma 1. Fix k. If there exists a neighborhood N of k such that $\sigma^*(k) = \sigma^*(k')$ for all $k' \in N$ then clearly $W(\sigma^*(k), k)$ is continuous at k.

To look at the kink points of W first notice the following basic property: $W(T, k)$ is decreasing in k and $W(R, k)$ is weakly increasing and both are continuous. Therefore, there exists k^* such that $\sigma^*(k) = T$ for all $k < k^*$ and $\sigma^*(k) = R$ for all $k \geq k^*$. But at k^* we have that $W(T, k^*) = W(R, k^*)$ so $W(\sigma^*(k), k)$ is continuous at k^* also. □

Proof of Proposition 4. The fact that $\mathcal{K}^R$ is non-empty is obvious (just consider arbitrarily large k).

For the second part of the proposition, suppose the DM has $\gamma = 0$ then $SC(R, k) = \psi(b) = SC(R, k')$ for all $k, k' \in \mathbb{R}_+$. This means that $W(R, k) = W(R, k') < W(T, 0)$ for all $k, k' \in \mathbb{R}_+$. But since $W(T, k)$ is decreasing in k it is the case that for any $k > 0$ we have that

$$\max\{W(T, k), W(R, k)\} < W(T, 0).$$

Thus $\mathcal{K}^* = \{0\}$. □

Proof of Proposition 5. To show this look at the utility of choosing R with stick contract of size k, this is simply

$$-\psi(b - e^{-\lambda} - e^{-\lambda t'} k)$$

if T is preferred to R by the SR self and 0 otherwise. k^* thus sets $SC(R, k^*) = 0$ which is exactly when

$$e^{-\lambda t'} k = b - e^{-\lambda}.$$

Algebra gives:

$$k^* = \frac{(b - e^{-\lambda})}{e^{-\lambda t'}}.$$

Corollary 1 follows from this argument as well. Note that the fraction has a denominator less than 1 which shrinks exponentially in t', thus k^* grows exponentially in t' and this shows Corollary 3. □

Proof of Corollary 2. Fix a DM by the argument in the proof of Proposition 1 there exists a $\underline{k}$ such that $k \geq \underline{k}$ implements resisting. By the indifference argument from the proof of Proposition 1 it must be that

$$W(0, P) > W(\underline{k}, I) = W(\underline{k}, P)$$

however by continuity it must be that for a small neighborhood $(\underline{k}, \overline{k})$ it is the case that

$$W(0, P) > W(k', P) \; \forall \; k' \in (\underline{k}, \overline{k}).$$

But each k' is also an element of $\mathcal{K}^R$ by construction, thus we have proved the corollary. □

Proof of Proposition 6. Note that for any r implemented at $t' > 1$ does not affect the preferences of the perfectly myopic SR self. By the way the carrot is financed, the LR self is also completely indifferent between the case where r is given at t' and lifetime consumption changes to finance the carrot or whether the carrot is not activated. In this case, no carrot can affect decisions. □

Proof of Proposition 7. To show this proposition, call the utility of the SR self from taking action x under a carrot of size r to be $u_{SR}(x, r)$. Then

$$u_{SR}(T, r) = b - e^{-\lambda}$$

while

$$u_{SR}(R, r) = e^{-\lambda t'} r.$$

This means that the self-control cost of R can be written as

$$\psi(b - e^{-\lambda} - e^{-\lambda t'} r)$$

while $u_{SR}(T, r) > u_{SR}(R, r)$ and 0 afterwards. This cost decreases continuously to 0 in r so that means there must exist an $\overline{r}$ such that for $r > \overline{r}$ the

$$b - 1 > -\psi(b - e^{-\lambda} - e^{-\lambda t'} r)$$

so the LR self chooses R. Thus

$$\mathcal{R}^R = \{r \in R \mid r \geq \overline{r}\}.$$

This also means that we can take

$$r^* = \frac{b - e^{-\lambda}}{e^{-\lambda t'}}$$

and we have that $r \geq r^*$ means the SR self weakly prefers R to T and thus that self-control costs are 0 and the DM chooses R. From this argument, combined with the one proving Proposition 6 we have that Corollary 4 is obvious. From the expression we can also see that Corollary 5 holds. □

Proof of Proposition 8. We can now consider the case of accepting a stick of size k^*. The SR self's utility of choosing k^* at time 0 is given by $u_{SR}(k^*, t = 0) = 0$ because the SR self correctly anticipates that R will be taken at time $t = 1$. Notice that this means that the SR's utility of choosing N at $t = 0$ is given by

$$u_{SR}(N, t = 0) = e^{-\lambda}(1 - e^{-\lambda})$$

which is greater than 0. Thus the self control cost from taking k^* from any menu that includes N is strictly positive. Note that an identical analysis applies to choosing a binding commitment when the option for no commitment is present.

Now consider the case of taking r^* at $t = 0$. Under a carrot of r^* the SR self is exactly indifferent between choosing R and T at $t = 1$. This means that

$$u_{SR}(c^*, t = 0) = e^{-\lambda}(1 - e^{-\lambda}) = u_{SR}(N, t = 0).$$

Thus the self control cost of implementing r^* at $t = 0$ is zero and so the DM strictly prefers carrots to sticks and binding commitments

Now, suppose that the agent gets a revision opportunity at some time $t \in (0, 1)$. If faced with a carrot, he has no incentive to revise. If faced with a stick or binding commitment, choosing N now carries an SR utility of

$$u_{SR}(N, t = t) = e^{-\lambda(1-t)}(1 - e^{-\lambda}) > u_{SR}(N, t = 0)$$

and thus a self control cost of $\psi(u_{SR}(N, t = t))$.

The ex-ante expected utility of taking a binding commitment or k^* at $t = 0$ given a Poisson arrival rate μ is continuously decreasing in μ and bounded above by $-(1 +$

$\mu)\psi(u_{SR}(N, t = 0))$ which decreases without bound. Thus there exists an $\overline{\mu}$ such that for $\mu > \overline{\mu}$ the DM prefers, at $t = 0$ to have no commitment than either k^* or a binding commitment. Note that because there is no temptation to revise a carrot, the expected welfare of a DM with a carrot contract is constant in μ. □

Proof of Proposition 9. Suppose that we set $k > \dfrac{b_{max}}{e^{-\lambda t'}}$ then for any realization of b the SR self prefers R to T so a stick can simulate binding commitment (thus the DM is at least indifferent to binding commitments). Now we show that this preference can be strict. Suppose that b can either be $\overline{b} < b_L < 1$ with probability p or $b_H > 1$ with probability $1 - p$. Thus the ex-ante expected utility of being uncommitted

$$p(b_L - 1) + (1 - p)(b_H - 1).$$

Now, suppose we take

$$k = \frac{b_L - e^{-\lambda}}{e^{-\lambda t'}}.$$

This means that at b_L the SR self is indifferent between R and T. The expected utility of taking k is then given by

$$p(0) + (1 - p)(b_H - k - 1).$$

Note that if b_H is very high then the DM is better off in this contract then under a binding commitment which delivers an expected utility of 0. Note also that given this problem we can set p to be close to 1 (in which case the ordering is stick $\succ$ binding $\succ$ no commitment) or close to 0 in which case the order is stick $\succ$ no commitment $\succ$ binding.

Suppose now that we set $r = \dfrac{1 - e^{-\lambda}}{e^{-\lambda t'}}$. This means that for all $b \leq 1$ the SR self prefers R to T. However, for all $b > 1$ the SR still prefers T to R so now LR and SR preferences are aligned in all cases. This means for any b the DM exerts no self-control and takes the choice that is LR optimal. Thus, carrots make the DM better off than either sticks or binding commitments. □

4.9 Appendix 4: Costs of Carrot Contracts

In the main text, carrots were funded by moving future consumption from future periods to time t'. We made the assumption that carrots were costless to make the comparison to costless sticks and binding commitments easier. We now discuss what the true cost of a carrot contract is: we will do this by bounding a DM's utility loss from perturbing a consumption stream in a particular way. We then calibrate this loss making an assumption of log utility and show that these intertemporal losses are quite reasonable and so even factoring this cost carrots remain a useful tool for DMs seeking commitment.

We look at the effects of taking a carrot on the welfare of the LR self. To do this, we simply consider a normal DM who lives in discrete time, discounts with rate δ and has a smooth, concave utility function u and has a flat consumption path $(c, c, c, c, \dots)$ until the end of time. Now, suppose that this DM is to move B units of consumption from the future into period $t = 1$. He does this in the following way: from each period $t = 2$ to $2 + N$ he borrowers $\frac{B}{N}$ units of consumption.

We now ask, what is the loss from $t = 2$ on from implementing this consumption path. This is exactly

$$\sum_{t=2}^{t+N+1} \delta^{t-1}(u(c) - u(c - \frac{B}{N})).$$

However, note that for a very patient DM and large N this loss can be well approximated by

$$Bu'(c).$$

Formally, the following is true:

Proposition 13. *For any $\epsilon > 0$ there exists $\delta < 1$ and N such that*

$$|\sum_{t=2}^{t+N+1} \delta^{t-1}(u(c) - u(c - \frac{B}{N})) - Bu'(c)| < \epsilon.$$

Proof. To show this note that we can use the fact that marginal utility is decreasing to bound this loss by

$$\sum_{t=2}^{t+N+1} \delta^{t-1}\frac{B}{N}u'(c - \frac{B}{N}).$$

This is because for each unit of consumption he loses, the DM loses at most $u'(c-\frac{B}{N})$ units of utility.

Now we can take N to be very large, and thus this is well approximated by

$$\sum_{t=2}^{t+N+1} \delta^{t-1}\frac{B}{N}u'(c).$$

Now we can also take δ to be very close to 1 in which case this is now well approximated by

$$N\frac{B}{N}u'(c)$$

which is exactly the result we want. □

Now, $Bu'(c)$ is an upper bound for the loss of a very patient LR self from $t = 2$ and on. There is an extra portion to be considered which is the fact that the DM gets to eat these B units at time $t = 1$. At $t = 1$ his gain is bounded below by $Bu'(c + B)$, again due to concavity. This means the total loss from the B perturbation of the DM's consumption stream is bounded above by

$$B(u'(c) - u'(c + B)).$$

We now consider whether this loss is relatively big or relatively small. To do this, we make a functional form assumption on the DM: we say that he has log utility. We also parametrize the size of B in terms of his daily consumption c so $B = rc$. For log utility, the LR loss *in utils* of the intertemporal disturbance is given by

$$rc(\frac{1}{c} - \frac{1}{(1+r)c}).$$

The expression above simplifies to

$$rc(\frac{rc}{c^2(1+r)})$$

which further simplifies to

$$\frac{r^2}{1+r}.$$

Now we can consider the relative cost of the intertemporal move B. Suppose that daily consumption c is $\$100$ per day, and that B is one day's consumption (a very nice dinner

compete with apertifs, paired wine and a delicious desert). This means that $r = 1$ and so the loss *in utils* of the move B is $\frac{1}{2}$. Note that daily utility is $log(100) \approx 4.6$ so the shadow cost (in utils) of a carrot of the same size as a full day's consumption is approximately 10% of a day's consumption in this case. The table below shows shadow intertemporal costs in terms of the utility of a day's consumption for various levels of B and c under the assumption of log utility. Note again that this is an upper bound on losses which may not always particularly tight because for extremely large values of B relative to c as our approximation of the gain at $t = 1$ will be off by quite a bit. However, for the values of c and B considered in the table the bound is relatively tight.

	$B = \$50$	$B = \$100$	$B = \$200$	$B = \$400$
$c = \$50$	12%	34%	81%	181%
$c = \$100$	3.6%	10%	29%	69%
$c = \$200$	$.9\%$	3.14%	9.4%	25%
$c = \$400$	$.2\%$	$.8\%$	2.7%	8.3%

Figure 4.3: Upper bound on shadow cost of a carrot of size B by levels of daily consumption as percentage of 1 day's consumption utility.

Thus, reasonably sized carrots (relative to daily consumption) are not that expensive to implement.

Chapter 5

Supporting Documentation

5.1 IRB and Human Subjects Approvals

All research involving human subjects in this thesis was approved by the IRB institution of Harvard University

For the first essay: experiments in the lab were approved by the Harvard Committee on the Use of Human Subjects (reference number F22190-101). Experiments using online recruitment reported in the same essay were approved under a separate protocol (reference number F17468-105).

For the second essay: all experiments were approved by the Committee on the Use of Human Subjects at the Harvard Business School (reference number F21910-101)

5.2 Funding

Funding for studies was provided by the Harvard University Economics Department, Harvard Business School and the LEAP Program at Harvard.

References

Ainslie, G. and N. Haslam, "Hyperbolic discounting," *Choice over time*, 1992, *5*, 57–92.

Ali, N., "Learning Self-Control," *Quarterly Journal of Economics*, 2011, *126* (2).

Amir, O., D.G. Rand, and Y.K. Gal, "Economic Games on the Internet: The Effect of $1 Stakes," *PloS one*, 2012, *7* (2), e31461.

Anderson, C.M. and L. Putterman, "Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism," *Games and Economic Behavior*, 2006, *54* (1), 1–24.

Andreoni, J., "Impure altruism and donations to public goods: a theory of warm-glow giving," *The Economic Journal*, 1990, *100* (401), 464–477.

_ , "An experimental test of the public-goods crowding-out hypothesis," *The American Economic Review*, 1993, pp. 1317–1327.

Ariely, D. and K. Wertenbroch, "Procrastination, deadlines, and performance: Self-control by precommitment," *Psychological Science*, 2002, *13* (3), 219.

Ashraf, N., D. Karlan, and W. Yin, "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines," *Quarterly Journal of Economics*, 2006, *121* (2), 635–672.

Ater, I., Y. Givati, and O. Rigbi, "Organizational Structure, Police Activity and Crime: Evidence from an Organizational Reform in Jails," *Working Paper*, 2012.

Axelrod, R., "An evolutionary approach to norms," *The American Political Science Review*, 1986, pp. 1095–1111.

Bardsley, N. and R. Sausgruber, "Conformity and reciprocity in public good provision," *Journal of Economic Psychology*, 2005, *26* (5), 664–681.

Baron, J. and I. Ritov, "Intuitions about penalties and compensation in the context of tort law," *Journal of Risk and Uncertainty*, 1993, *7* (1), 17–33.

_ **and** _ , "The role of probability of detection in judgments of punishment," *Journal of Legal Analysis*, 2009, *1* (2), 553–590.

Becker, Gary S., "Crime and punishment: An economic approach," *Journal of Political Economy*, 1968, *76* (2), 169–217.

Becker, GM, MH DeGroot, and J. Marschak, "Measuring utility by a single-response sequential method.," *Behavioral science*, 1964, *9* (3), 226.

Benhabib, J. and A. Bisin, "Modeling internal commitment mechanisms and self-control: A neuroeconomics approach to consumption-saving decisions," *Games and Economic Behavior*, 2005, *52* (2), 460–492.

Berdejo, C. and N. Yuchtman, "Crime, Punishment and Politics: An Analysis of Political Cycles in Criminal Sentencing," *Unpublished manuscript, Harvard University*, 2009.

Borghans, L., B.H.H. Golsteyn, J.J. Heckman, and H. Meijers, "Gender differences in risk aversion and ambiguity aversion," Technical Report, National Bureau of Economic Research 2009.

Bryan, G., D. Karlan, and S. Nelson, "Commitment devices," *Annual Review of Economics*, 2010, *2* (1).

Bushway, S., M.A. Stoll, and D.F. Weiman, *Barriers to Reentry?: The Labor Market for Released Prisoners in Post-industrial America*, Russell Sage Foundation Publications, 2007.

Camerer, C. and M. Weber, "Recent developments in modeling preferences: Uncertainty and ambiguity," *Journal of risk and uncertainty*, 1992, *5* (4), 325–370.

Camerer, C.F., T.H. Ho, and J.K. Chong, "A cognitive hierarchy model of games," *The Quarterly Journal of Economics*, 2004, *119* (3), 861–898.

Carlsmith, K.M., J.M. Darley, and P.H. Robinson, "Why do we punish?: Deterrence and just deserts as motives for punishment.," *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology*, 2002, *83* (2), 284.

Coffman, L.C., "Intermediation reduces punishment (and reward)," *American Economic Journal: Microeconomics*, 2011, *3* (4), 77–106.

Cornes, R. and T. Sandler, "The comparative static properties of the impure public good model," *Journal of Public Economics*, 1994, *54* (3), 403–421.

Costa-Gomes, M., V.P. Crawford, and B. Broseta, "Cognition and Behavior in Normal-Form Games: An Experimental Study," *Econometrica*, 2003, *69* (5), 1193–1235.

Cushman, F., A. Dreber, Y. Wang, and J. Costa, "Accidental outcomes guide punishment in a ?trembling hand? game," *PloS one*, 2009, *4* (8), e6699.

Danziger, S., J. Levav, and L. Avnaim-Pesso, "Extraneous factors in judicial decisions," *Proceedings of the National Academy of Sciences*, 2011, *108* (17), 6889–6892.

Dekel, E., B.L. Lipman, and A. Rustichini, "Temptation-driven preferences," *Review of Economic Studies*, 2009, *76* (3), 937–971.

DellaVigna, S. and U. Malmendier, "Paying not to go to the gym," *American Economic Review*, 2006, *96* (3), 694–719.

Eckel, C.C. and P.J. Grossman, "Men, women and risk aversion: Experimental evidence," *Handbook of experimental economics results*, 2008, *1*, 1061–1073.

Ellsberg, D., "Risk, ambiguity, and the Savage axioms," *The Quarterly Journal of Economics*, 1961, pp. 643–669.

Epstein, L.G. and M. Schneider, "Ambiguity, information quality, and asset pricing," *The Journal of Finance*, 2008, *63* (1), 197–228.

Fehr, E. and K.M. Schmidt, "A theory of fairness, competition, and cooperation," *The Quarterly Journal of Economics*, 1999, *114* (3), 817–868.

_ **and U. Fischbacher**, "Third-party punishment and social norms," *Evolution and human behavior*, 2004, *25* (2), 63–87.

Fischbacher, U., "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 2007, *10* (2), 171–178.

Fox, C.R. and A. Tversky, "Ambiguity aversion and comparative ignorance," *The quarterly journal of economics*, 1995, *110* (3), 585–603.

Fudenberg, D. and D.K. Levine, "A dual-self model of impulse control," *The American Economic Review*, 2006, *96* (5), 1449–1476.

_ **and** _ , "Timing and Self-Control," *Econometrica*, 2012, *80* (1), 1–42.

_ **and P.A. Pathak**, "Unobserved punishment supports cooperation," *Journal of Public Economics*, 2010, *94* (1-2), 78–86.

Garland, D., *The culture of control: Crime and social order in contemporary society*, Oxford University Press US, 2001.

Gintis, H., S. Bowles, R.T. Boyd, and E. Fehr, *Moral sentiments and material interests: The foundations of cooperation in economic life*, Vol. 6, MIT press, 2005.

Glazer, A. and K.A. Konrad, "A signaling explanation for charity," *The American Economic Review*, 1996, *86* (4), 1019–1028.

Greene, J. and J. Haidt, "How (and where) does moral judgment work?," *Trends in cognitive sciences*, 2002, *6* (12), 517–523.

Gul, F. and W. Pesendorfer, "Temptation and self-control," *Econometrica*, 2001, *69* (6), 1403–1435.

_ **and** _ , "Self-control and the theory of consumption," *Econometrica*, 2004, *72* (1), 119–158.

Haidt, J., "The emotional dog and its rational tail: a social intuitionist approach to moral judgment.," *Psychological Review; Psychological Review*, 2001, *108* (4), 814.

Halevy, Y., "Ellsberg revisited: An experimental study," *Econometrica*, 2007, *75* (2), 503–536.

Hare, T.A., C.F. Camerer, and A. Rangel, "Self-control in decision-making involves modulation of the vmPFC valuation system," *Science*, 2009, *324* (5927), 646.

Horton, J.J., D.G. Rand, and R.J. Zeckhauser, "The online laboratory: Conducting experiments in a real labor market," *Experimental Economics*, 2011, *14* (3), 399–425.

Houser, D., D. Schunk, J.K. Winter, and E. Xiao, "Temptation and Commitment in the Laboratory," *Institute for Empirical Research in Economics*, 2010.

Hsu, M., M. Bhatt, R. Adolphs, D. Tranel, and C.F. Camerer, "Neural systems responding to degrees of uncertainty in human decision-making," *Science*, 2005, *310* (5754), 1680–1683.

Huettel, S.A., C.J. Stowe, E.M. Gordon, B.T. Warner, and M.L. Platt, "Neural signatures of economic preferences for risk and ambiguity," *Neuron*, 2006, *49* (5), 765–775.

Kahneman, D., "Maps of bounded rationality: Psychology for behavioral economics," *American economic review*, 2003, *93* (5), 1449–1475.

_ **and A. Tversky**, *Choices, values, and frames*, Cambridge University Press, 2000.

_ **, P.P. Wakker, and R. Sarin**, "Back to Bentham? Explorations of experienced utility," *The Quarterly Journal of Economics*, 1997, *112* (2), 375–406.

Kaur, S., M. Kremer, and S. Mullainathan, "Self-control and the development of work arrangements," in "American Economic Review Papers and Proceedings" 2010.

Knight, F.H., *Risk, uncertainty and profit*, Dover publications, 1921.

Laibson, D., "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 1997, *112* (2), 443–477.

Loewenstein, G. and R.H. Thaler, "Anomalies: intertemporal choice," *The journal of economic perspectives*, 1989, *3* (4), 181–193.

McClure, S.M., K.M. Ericson, D.I. Laibson, G. Loewenstein, and J.D. Cohen, "Time discounting for primary rewards," *Journal of Neuroscience*, 2007, *27* (21), 5796.

Muraven, M. and R.F. Baumeister, "Self-regulation and depletion of limited resources: Does self-control resemble a muscle?," *Psychological Bulletin*, 2000, *126* (2), 247–259.

Noor, J. and N. Takeoka, "Uphill self-control," *Theoretical Economics*, 2010, *5* (2), 127–158.

Ostrom, E., J. Walker, and R. Gardner, "Covenants with and without a sword: Self-governance is possible," *The American Political Science Review*, 1992, pp. 404–417.

Pager, D., *Marked: Race, crime, and finding work in an era of mass incarceration*, University of Chicago Press, 2007.

Paolacci, G., J. Chandler, and P.G. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision Making*, 2010, *5* (5), 411–419.

Peysakhovich, A. and D. Rand, "Building cooperative norms," *Working paper*, 2012.

Polinsky, A.M. and S. Shavell, "The fairness of sanctions: some implications for optimal enforcement policy," *American Law and Economics Review*, 2000, *2* (2), 223–237.

Posner, R.A., *How judges think*, Harvard University Press, 2008.

Quervain, D.J.F. De, U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, and E. Fehr, "The neural basis of altruistic punishment.," *Science; Science*, 2004.

Rabin, M., "Incorporating fairness into game theory and economics," *The American Economic Review*, 1993, pp. 1281–1302.

Rand, DG, J.D. Greene, and M.A Nowak, "Spontaneous giving and calculated greed," *Nature*, in Press.

Rasmusen, E.B., "Stigma and self-fulfilling expectations of criminality," *Journal of Law and Economics*, 1996, *39*, 519–544.

Ritov, I. and J. Baron, "Reluctance to vaccinate: Omission bias and ambiguity," *Journal of Behavioral Decision Making*, 1990, *3* (4), 263–277.

Roth, A.E., "The economist as engineer: Game theory, experimentation, and computation as tools for design economics," *Econometrica*, 2003, *70* (4), 1341–1378.

_ **and J.H. Kagel**, *The handbook of experimental economics*, Vol. 1, Princeton University Press Princeton, 1995.

Shavell, S., "A model of optimal incapacitation," *The American Economic Review*, 1987, *77* (2), 107–110.

Shiv, B. and A. Fedorikhin, "Heart and mind in conflict: The interplay of affect and cognition in consumer decision making," *Journal of Consumer Research*, 1999, *26* (3), 278–292.

Singer, T., B. Seymour, J.P. O'Doherty, K.E. Stephan, R.J. Dolan, and C.D. Frith, "Empathic neural responses are modulated by the perceived fairness of others," *Nature*, 2006, *439* (7075), 466–469.

Sunstein, C.R., D. Schkade, and D. Kahneman, "Do people want optimal deterrence," *J. Legal Stud.*, 2000, *29*, 237.

Tymula, A., L.A.R. Belmaker, A.K. Roy, L. Ruderman, K. Manson, P.W. Glimcher, and I. Levy, "Adolescents risk-taking behavior is driven by tolerance to ambiguity," *Proceedings of the National Academy of Sciences*, 2012, *109* (42), 17135–17140.